PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-117937
(43)Date of publication of application : 27.04.2001

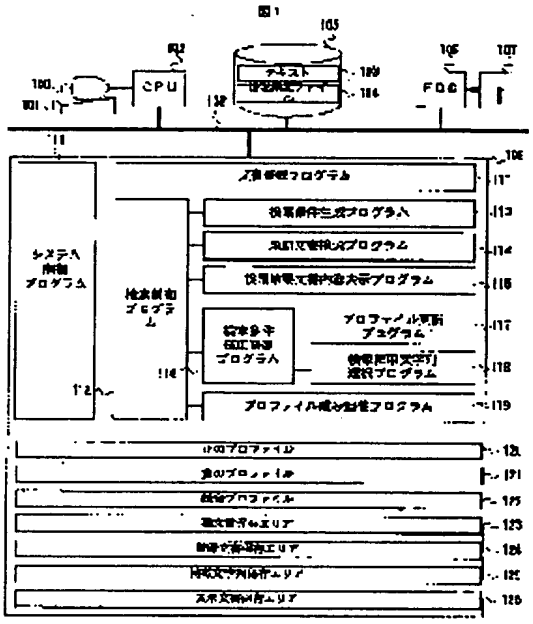
(51)Int.Cl. G06F 17/30

(21)Application number : 11-297604
(22)Date of filing : 20.10.1999

(71)Applicant : HITACHI LTD
(72)Inventor : INABA YASUHIKO
TADA KATSUMI
SUGAYA NATSUKO
MATSUBAYASHI TADATAKA
YAMAGUCHI AKIHIKO
KAWASHITA YASUSHI

(54) METHOD AND DEVICE FOR RETRIEVING DOCUMENT

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a system capable of easily improving the accuracy of retrieval on the basis of a suitable or unsuitable user evaluation to the retrieved result in the case of similar document retrieval for retrieving the document of contents similar to a document designated by a user.
SOLUTION: Retrieval condition data are updated while using a character string extracted from an evaluation object document, and retrieval is performed while using a character string, which is not contained in one part or all character strings extracted from a document evaluated as desired one, between a character string extracted from a document evaluated desired for the user and a character string extracted from a document evaluated non-desired for the user.



LEGAL STATUS

[Date of request for examination] 19.03.2003

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2001－117937
（P2001－117937A）
(43)公開日　平成13年４月27日(2001.4.27)

(51)Int.Cl.[7]	識別記号	Ｆ Ｉ		テーマコード（参考）
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/403	３４０Ｂ	５Ｂ０７５
		15/40	３７０Ａ	
		15/403	３５０Ｃ	

審査請求　未請求　請求項の数６　ＯＬ　（全 25 頁）

(21)出願番号　特願平11－297604

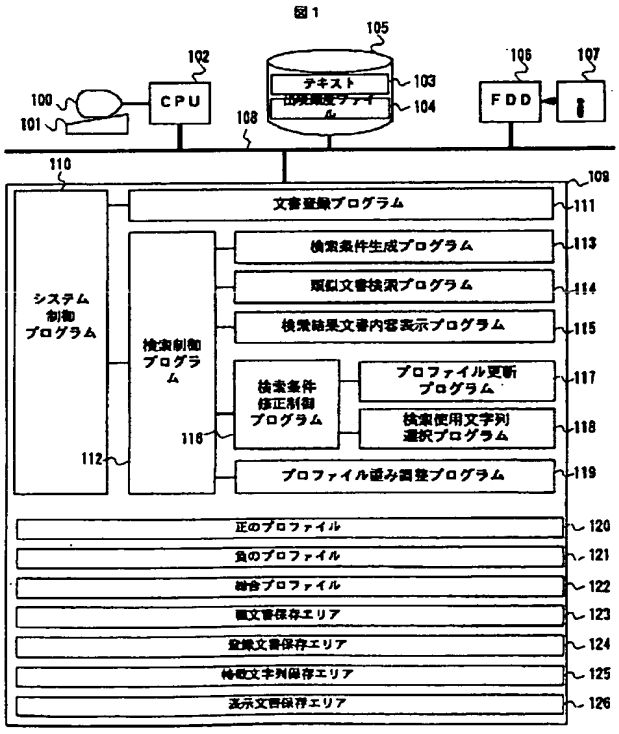
(22)出願日　平成11年10月20日(1999.10.20)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　稲場　靖彦
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所システム開発本部内
(72)発明者　多田　勝己
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所システム開発本部内
(74)代理人　100075096
弁理士　作田　康夫

最終頁に続く

(54)【発明の名称】　文書検索方法および装置

(57)【要約】
【課題】ユーザが指定した文書と内容的に類似した文書を検索する類似文書検索において、検索結果に対する適、不適といったユーザ評価にもとづいて、検索の精度を簡易に向上させることのできるシステムを提供する。
【解決手段】評価対象文書から抽出された文字列を用いて検索条件データを更新し、ユーザが所望のものであると評価をした文書から抽出した文字列と、ユーザが所望のものでないと評価した文書から抽出した文字列のうち所望のものであると評価をした文書から抽出した文字列の一部または全部に含まれないもの、を用いて検索を行なう。



【特許請求の範囲】
【請求項１】文字列に付与された重みを含む検索条件により文書データベースを検索し、該検索により得られた文書に対してユーザが入力した「所望である」または「所望でない」の評価を受け取り、上記検索の結果得られた文書から抽出した文字列の重みを上記評価に基づき変更して検索する文書検索方法において、
上記「所望である」と評価した文書から抽出した第一の文字列に正の重みを付与し、
上記「所望でない」と評価した文書から抽出した第二の文字列に負の重みを付与し、
上記第二の文字列のうち上記第一の文字列と一致するとともに当該第一の文字列の重みが所定値以上となるものを除外したものとその重みおよび上記第一の文字列とその重みとを含む検索条件を生成して検索することを特徴とした文書検索方法。
【請求項２】文字列に付与された重みを含む検索条件により文書データベースを検索し、該検索により得られた文書に対してユーザが入力した「所望である」または「所望でない」の評価を受け取り、上記検索の結果得られた文書から抽出した文字列の重みを上記評価に基づき変更して検索する文書検索方法において、
上記「所望である」と評価した文書から第一の文字列を抽出し、
上記「所望でない」と評価した文書から抽出した文字列で上記第一の文字列と一致する場合は、当該第一の文字列の重みが所定値以下の場合は上記抽出した文字列を第二の文字列として抽出し、第二の文字列の重みを第一の文字列の重みよりも低くし、一致しない場合は上記抽出した文字列を第二の文字列として抽出し、第二の文字列の重みを第一の文字列の重みよりも低くすることを特徴とした文書検索方法。
【請求項３】文字列に付与された重みを含む検索条件により文書データベースを検索し、該検索により得られた文書に対してユーザが入力した「所望である」または「所望でない」の評価を受け取り、上記検索の結果得られた文書から抽出した文字列の重みを上記評価に基づき変更して検索する文書検索方法において、
上記「所望である」と評価した文書から第一の文字列を抽出し、
上記「所望でない」と評価した文書から抽出した文字列で上記第一の文字列と一致しない場合は上記抽出した文字列を第二の文字列として抽出し、第二の文字列の重みを第一の文字列の重みよりも低くすることを特徴とした文書検索方法。
【請求項４】請求項１または請求項２記載の文書検索方法において、
上記検索結果文書に対するユーザの評価は、ひとつ以上の段階評価を設定する方法であり、
文字列の重みの変更方法は、前記評価に応じて多段階に変更の程度を設定する方法であることを特徴とする文書検索方法。
【請求項５】請求項１または請求項２記載の文書検索方法において、ユーザの評価に基づき検索条件を更新する際に、
ユーザが評価した文書から抽出した文字列について、検索条件に用いるか否かをユーザが選択情報を入力して選択することすることを特徴とした文書検索方法。
【請求項６】文字列に付与された重みを含む検索条件により文書データベースを検索し、該検索により得られた文書に対してユーザが入力した「所望である」または「所望でない」の評価を受け取り、上記検索の結果得られた文書から抽出した文字列の重みを上記評価に基づき変更して検索する文書検索装置において、
上記「所望である」と評価した文書から抽出した第一の文字列に正の重みを付与し、
上記「所望でない」と評価した文書から抽出した文字列が上記第一の文字列と一致する場合は当該第一の文字列の重みが所定値以下の場合は上記抽出した文字列を第二の文字列として負の重みを付与し、上記第一の文字列と一致しない場合は上記抽出した文字列を第二の文字列として負の重みを付与する手段と、
上記第一の文字列とその重みおよび上記第二の文字列とその重みとを含む検索条件を生成して検索する手段とを備えたことを特徴とした文書検索方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、検索条件に基づいて文書データベースから文書を検索する方法および装置に関し、その検索の結果として得られた文書に対してユーザが評価を与え、その評価に基づき検索条件を変更する方法および装置に関する。
【０００２】
【従来の技術】近年、パーソナルコンピュータやインターネット等の普及に伴い、電子化文書が急激に増加している。このような状況において、ユーザが所望する情報を含んだ文書を高速かつ効率的に検索したいという要求が高まってきている。
【０００３】このような要求に応えるための検索技術としてレリバンスフィードバックとよばれる技術がある。この技術は、全文検索や類似文書検索による検索結果に対して、ユーザが「所望の文書である」か「所望の文書でない」かなどの評価をシステムに入力し、その評価情報を検索条件に反映させることにより、その後の検索結果を改善する技術である。
【０００４】具体的な処理の内容としては、例えば「"Information Retrieval",William B.Frakes / Rocardo Baeza-Yates, Prentice Hall PTR, 1992 p.p.241～263」に示されるように、ユーザが所望であると評価した文書から抽出した単語に関する検索条件中の重みを加算し、

所望でないと評価した文書から抽出された単語に関する検索条件中の重みを減算する方法がある。以下この技術を従来技術１と呼ぶ。検索条件中のある単語について、具体的な重みの加減算の方法の例を式１に示す。

【０００５】

【数１】

$$W' = W + \alpha \sum_{i}^{P} FP(i) - \beta \sum_{j}^{N} FN(j) \qquad \cdots(数1)$$

【０００６】ここでW'はその単語の新たな重み、Wは元の重みであり、ＦＰ（ｉ）は所望であると評価されたｉ番目の文書におけるその単語の出現回数、ＦＮ（ｊ）は所望でないと評価されたｊ番目の文書におけるその単語の出現回数である。また、Ｐは所望であると評価された文書の数、Ｎは所望でないと評価された文書の数である。なお、α、βはパラメータである。ここで、この新たな重みW'は負になってもよく、そのような場合は、その単語が含まれる文書は類似度が下がることになる。

【０００７】この従来技術１によるレリバンスフィードバック処理の例を図２に示す。本図に示す例は、ユーザが「高校野球」に関する文書を所望する場合に、「サッカーに続き高校野球が開幕した」という文書を種文書に選んだ場合である。その後、「サッカー」に関するノイズ文書に対し「所望でない」と評価をして、システムに入力した場合である。この結果、本図に示すように「サッカー」という単語の重みが下がり、以後「サッカー」に関する文書の類似度を下げることができる。

【０００８】

【発明が解決しようとする課題】しかし、従来技術１による方式では、ユーザが「所望のものでない」といった評価をしたときに検索結果が改善しない場合がある。この問題を図３を用いて説明する。本図に示した例は、「高校野球」に関する文書を所望する場合に、「高校サッカーが開幕した・・・」といったノイズ文書に対し「所望の文書でない」と評価した場合である。このとき従来技術１によれば、このノイズ文書から「高校」「サッカー」「開幕」といった単語を抽出し、検索条件中のそれぞれの単語の重みを減算することになる。この場合、「サッカー」の重みを減算するだけでなく、「高校」という単語の重みまでも減算してしまう。その結果、更新された検索条件によって検索を行なうと、「高校野球」に関する文書の類似度が、「プロ野球」「社会人野球」といった文書の類似度よりも低くなってしまうという問題がある。

【０００９】このように、従来の方法によりユーザが「所望のものでない」と評価した文書から抽出した単語の重みを単純に減算すると、ユーザが所望とする概念を表す単語の重みまで減算してしまい、検索結果が改善しないという問題がある。

【００１０】本発明の目的は、ユーザが「所望のものでない」といった評価を与えた文書から抽出した情報のうち適切なものを使用して、検索結果を改善することにある。

【００１１】

【課題を解決するための手段】上記課題を解決するため、第１の手段として、文字列に付与された重みを含む検索条件により文書データベースを検索し、該検索により得られた文書に対してユーザが入力した「所望である」または「所望でない」の評価を受け取り、上記検索の結果得られた文書から抽出した文字列の重みを上記評価に基づき変更して検索する文書検索方法において、上記「所望である」と評価した文書から抽出した第一の文字列に正の重みを付与し、上記「所望でない」と評価した文書から抽出した第二の文字列に負の重みを付与し、第二の文字列のうち上記第一の文字列と一致するもとともに当該第一の文字列の重みが所定値以上ものを除外したものとその重みおよび上記第一の文字列とその重みとを含む検索条件を生成して検索する。

【００１２】この方法により、ユーザが所望のものと評価した文書から抽出した所望の内容を特徴付ける文字列に付与された負の重みにより検索精度を下げてしまうという課題を改善することができる。

【００１３】また、第２の手段は、文字列に付与された重みを含む検索条件により文書データベースを検索し、該検索により得られた文書に対してユーザが入力した「所望である」または「所望でない」の評価を受け取り、上記検索の結果得られた文書から抽出した文字列の重みを上記評価に基づき変更して検索する文書検索方法において、上記「所望である」と評価した文書から第一の文字列を抽出し、上記「所望でない」と評価した文書から抽出した文字列で上記第一の文字列と一致する場合は、当該第一の文字列の重みが所定値以下の場合は上記抽出した文字列を第二の文字列として抽出し、第二の文字列の重みを第一の文字列の重みよりも低くし、一致しない場合は上記抽出した文字列を第二の文字列として抽出し、第二の文字列の重みを第一の文字列の重みよりも低くする。

【００１４】この方法により、ユーザが所望のものと評価した文書から抽出した所望の内容を特徴付ける文字列に、負の重みを付与してしまい以降の検索精度を下げてしまうという課題を改善できる。

【００１５】

【発明の実施の形態】以下、本発明の第一の実施例について説明する。

【００１６】まず、本発明の第一の実施例のシステム構成を図１に示す。本実施例におけるシステムは、ディスプレイ１００、キーボード１０１、中央演算処理装置（ＣＰＵ）１０２、磁気ディスク装置１０５、フロッピディスクドライブ（ＦＤＤ）１０６、主メモリ１０９およびこれらを結ぶバス１０８から構成される。

【００１７】磁気ディスク装置１０５は二次記憶装置の

一つであり、テキスト１０３、出現頻度ファイル１０４が格納される。ＦＤＤ１０６を介してフロッピディスク１０７に格納されている情報が、主メモリ１０９あるいは磁気ディスク装置１０５へ読み込まれる。

【００１８】主メモリ１０９には、システム制御プログラム１１０、文書登録プログラム１１１、検索制御プログラム１１２が格納される。検索制御プログラム１１２は、検索条件生成プログラム１１３、類似文書検索プログラム１１４、検索結果文書内容表示プログラム１１５、検索条件修正制御プログラム１１６、およびプロファイル重み調整プログラム１１９で構成される。ここで、検索条件修正制御プログラム１１６は、プロファイル更新プログラム１１７、および検索使用文字列選択プログラム１１８で構成される。

【００１９】また、正のプロファイル１２０、負のプロファイル１２１、総合プロファイル１２２、種文書保存エリア１２３、登録文書保存エリア１２４、特徴文字列保存エリア１２５、および表示用文書保存エリア１２６が同じく主メモリ１０９に確保される。

【００２０】ここで、正のプロファイル１２０、負のプロファイル１２１、総合プロファイル１２２とは後述する図１５に示すように、いずれも幾つかの検索文字列とその重みを保持したデータである。正のプロファイル１２０には、ユーザが所望であると評価した文書から抽出した文字列が格納される。負のプロファイル１２１には、ユーザが所望のものでないと評価した文書から抽出した文字列が格納される。総合プロファイル１２２は、正負のプロファイルから選択された検索に用いる文字列が格納される。

【００２１】以下に、第一の実施例における、各プログラムの処理手順について説明する。

【００２２】まず、システム制御プログラム１１０の処理手順について図４のＰＡＤ（Ｐｒｏｂｌｅｍ Ａｎａｌｙｓｉｓ Ｄｉａｇｒａｍ）図を用いて説明する。

【００２３】システム制御プログラム１１０は、まずステップ４０１においてユーザがキーボードから入力したコマンドを解析する。

【００２４】次にステップ４０２において、このコマンドが文書登録のコマンドであると解析された場合には、ステップ４０４で文書登録プログラム１１１を起動して文書の登録を行なう。

【００２５】またステップ４０３において、検索実行のコマンドであると解析された場合には、ステップ４０５で検索制御プログラム１１２を起動して文書の検索を行なう。

【００２６】以上が、システム制御プログラム１１０の処理手順である。

【００２７】次に、図４に示したステップ４０４でシステム制御プログラムにより起動される、文書登録プログラム１１１について図５のＰＡＤ図を用いて説明する。

【００２８】文書登録プログラム１１１は、まずステップ５０１においてＤ１０６に挿入されたフロッピディスク１０７から登録すべき文書データを読み込み、これをテキスト１０３として磁気ディスク装置１０５に格納する。文書データは、フロッピディスク１０７を用いて入力するだけに限らず、通信回線やＣＤ－ＲＯＭ装置（図１には示していない）等を用いて他の装置から入力するような構成を取ることも可能である。

【００２９】次にステップ５０２で、検索対象文書から抽出される自立語の可能性がある文字列（以下、特徴文字列と呼ぶ）がどの文書に何回出現したかを高速に抽出するためのデータとして、出現頻度ファイル１０４を各登録対象文書について生成する。ここで出現頻度ファイルの生成方法としては「特開平１１－１４３９０２号広報」に開示されている出現頻度ファイルの生成方法と同一の方法でも良いし、形態素解析等を用いて各文書中の単語を抽出する方法やニューラルネットワークの学習データを用いた方法でもかまわない。また、単純ｎ－ｇｒａｍを抽出する方法であってもかまわない。

【００３０】以上が、文書登録プログラム１１１の処理手順である。次に、図４に示したステップ４０５でシステム制御プログラムにより起動される、検索制御プログラム１１２の処理手順を図６のＰＡＤ図を用いて説明する。

【００３１】検索制御プログラム１１２は、まずステップ６０１において検索条件生成プログラム１１３を起動し、検索条件を生成する。

【００３２】次にステップ６０２において、ステップ６０３～ステップ６１２の処理を、ステップ６０４においてユーザから検索セッションの終了が要求されたと解析されるまで繰り返す。

【００３３】この繰り返し処理では、まずステップ６０３において、類似文書検索プログラム１１４を起動し、ステップ６０１で生成された検索条件にもとづき類似文書検索を行なう。

【００３４】次にステップ６０４において、キーボードから入力されるコマンドを解析する。

【００３５】次にステップ６０５において、このコマンドが文書の内容表示コマンドであると解析された場合には、ステップ６０９で検索結果文書内容表示プログラム１１５を起動し、指定された検索結果文書の内容を表示する。

【００３６】次にステップ６０６において、検索結果文書に対するユーザの評価の入力コマンドであると解析された場合には、ステップ６１０で検索条件修正制御プログラム１１６を起動し、検索条件を修正する。

【００３７】次にステップ６０７において、プロファイルの内容調整コマンドであると解析された場合には、ステップ６１１でプロファイル重み調整プログラム１１９を起動し、プロファイルの内容を調整する。

【００３８】次にステップ６０８において、検索セッション終了コマンドであると解析された場合には、ステップ６１２で、正のプロファイル１２０、負のプロファイル１２１、および総合プロファイル１２２の内容をクリアし、ステップ６０２の繰り返しを終了する。

【００３９】以上が検索制御プログラム１１２の処理手順である。

【００４０】次に、図６に示したステップ６０１で検索制御プログラムにより起動される、検索条件生成プログラム１１３の処理手順を図７のＰＡＤ図を用いて説明する。

【００４１】検索条件生成プログラム１１３は、まずステップ７０１において、キーボード１０１から入力される種文書を読み込み、種文書保存エリア１２３に格納する。

【００４２】次にステップ７０２において、種文書保存エリア１２３に格納された種文書から特徴文字列を抽出し、種文書内出現回数を計数して、特徴文字列保存エリア１２５に格納する。

【００４３】ここで、特徴文字列を抽出する方法は、図５に示した文書登録プログラム１１１のステップ５０２における方法を用いても良いし、その他の方法を用いても良い。

【００４４】次にステップ７０３において、ステップ７０２で抽出した特徴文字列をステップ７０２で計数した出現回数と共に総合プロファイル１２２に書き込む。ここで総合プロファイル１２２は、後述する図１５に示すように特徴文字列とその重みが保持されたものであり、後述するように類似文書検索プログラム１１４の入力として使用する。ここで重みとしては種文書内出現回数を用いるものとするが、他のものを用いても良い。また、ここで総合プロファイル１２２に書き込む文字列は、ステップ７０２で抽出した特徴文字列のうち重みの上位から所定数のものに限定しても良い。

【００４５】次にステップ７０４において、ステップ７０２で抽出した文字列をステップ７０２で計数した出現回数と共に正のプロファイル１２０に書き込む。この正のプロファイル１２０は、後述するように、検索結果文書に対しユーザが評価をした場合に、検索条件を修正する際に使用する。また、ここで正のプロファイル１２０に書き込む文字列は、ステップ７０２で抽出した特徴文字列のうち重みの上位のもの所定数に限定しても良い。

【００４６】以上が、検索条件生成プログラム１１３の処理手順である。

【００４７】次に、図６に示したステップ６０３で検索制御プログラムにより起動される、類似文書検索プログラム１１４の処理手順を図８のＰＡＤ図を用いて説明する。

【００４８】類似文書検索プログラム１１４は、まずステップ８０１において、図７に示したステップ７０３で検索条件生成プログラム１１３により生成された総合プロファイル１２２を読み込む。

【００４９】次にステップ８０２において、出現頻度ファイル１０４を読み込む。

【００５０】次にステップ８０３において、総合プロファイル１２２内の特徴文字列の重みと、出現頻度ファイル１０４内の各文書における該文字列の出現頻度から、テキスト１０３内の各文書の類似度を算出する。ここで類似度の算出式としては、例えば以下の式２のようなものを用いる。

【００５１】

【数２】

$$S(D) = \sum_{i}^{N} \{ Frq(i) \times w(i) \} \quad \cdots (数2)$$

【００５２】この式で、Ｓ（Ｄ）はテキスト１０３内の文書番号Ｄの類似度であり、Ｆｒｑ（ｉ）は出現頻度ファイル１０４内の単語ｉの文書Ｄにおける出現頻度であり、ｗ（ｉ）は総合プロファイル内の単語ｉの重みである。ここで類似度算出式としては、これ以外のものを用いても構わない。

【００５３】次にステップ８０４において、テキスト１０３内の各文書の文書番号を類似度の順に降順にソートし、ディスプレイ１００に出力する。ここで、類似度の上位所定件のみを出力するようにしても良いし、所定の類似度を上回るもののみを出力するようにしても良い。また、文書にタイトルのような属性があればそれを出力しても良い。

【００５４】以上が、類似文書検索プログラム１１４の処理手順である。

【００５５】次に、図６に示したステップ６０９で検索制御プログラムにより起動される、検索結果文書内容表示プログラム１１５の処理手順を図９のＰＡＤ図を用いて説明する。

【００５６】検索結果文書内容表示プログラム１１５は、まずステップ９０１において、ユーザがキーボード１０１から入力する文書番号を読み込む。

【００５７】次にステップ９０２において、ステップ９０１で入力された文書番号に該当する文書を登録文書保存エリア１２４に読み込む。

【００５８】次にステップ９０３において、ステップ９０４で該文書を最後まで読み込むまで以下に示すステップ９０４からステップ９０７の処理を繰り返す。

【００５９】ステップ９０３の繰り返し処理では、まずステップ９０４において、登録文書保存エリア１２４の文書の文字列を順次読み込み、総合プロファイル１２２に格納された文字列と照合する。

【００６０】次にステップ９０５において、ステップ９０４で読み込んだ文字列が総合プロファイル１２２において正の重みを持つ文字列と一致した場合には、ステップ９０８で「該文字列を赤色表示する」という情報を付

与して表示用文書保存エリア１２６に追加する。ここで例えばＨＴＭＬ（ＨｙｐｅｒＴｅｘｔ ＭａｒｋｕｐＬａｎｇｕａｇｅ）の形式で表示する場合は、該文字列の前後に赤色表示を表すタグを挿入し、表示用文書保存エリア１２６に追加する。ここで、重みが所定値以下の文字列や、重みの上位所定件に含まれないものは、この処理の対象外にするなどしても構わない。また、表示色は別の色を用いても構わない。

【００６１】次にステップ９０６において、ステップ９０４で読み込んだ文字列が総合プロファイル１２２において負の重みを持つ文字列と一致した場合には、ステップ９０９で「該文字列を青色表示する」という情報を付与して表示用文書保存エリア１２６に追加する。ここで例えばＨＴＭＬの形式で表示する場合は、該文字列の前後に青色表示を表すタグを挿入し、表示用文書保存エリア１２６に追加する。ここで、重みが所定値以下の文字列や、重みの上位所定件に含まれないものは、この処理の対象外にするなどしても構わない。また、表示色はステップ９０８で指定する色以外の別の色を用いても構わない。

【００６２】次にステップ９０７において、ステップ９０４で読み込んだ文字列が総合プロファイル内の文字列と一致しない場合には、ステップ９１０で「該文字列を黒色表示する」という情報を付与して表示用文書保存エリア１２６に追加する。ここで例えばＨＴＭＬの形式で表示する場合は、該文字列の前後に黒色表示を表すタグを挿入し、表示用文書保存エリア１２６に追加する。ここで、表示色はステップ９０８、９０９で指定する以外の別の色を用いても構わない。

【００６３】次にステップ９１１において、表示用文書保存エリア１２６に保存された内容をディスプレイ１００に表示する。

【００６４】以上が、検索結果文書内容表示プログラム１１５の処理手順である。

【００６５】次に、図６に示したステップ６１０で検索制御プログラムにより起動される、検索条件修正制御プログラム１１６の処理手順を図１０のＰＡＤ図を用いて説明する。

【００６６】検索条件修正制御プログラム１１６は、まずステップ１００１においてプロファイル更新プログラム１１７を起動し、正のプロファイル１２０および負のプロファイル１２１の内容を更新する。

【００６７】次にステップ１００２において、検索使用文字列選択プログラム１１８を起動し、ステップ１００１で更新された正のプロファイル１２０および負のプロファイル１２１の内容にもとづき、総合プロファイル１２２の内容を更新する。

【００６８】以上が検索条件修正プログラム１１６の処理手順である。

【００６９】次に、図６に示したステップ６１１で検索制御プログラムにより起動される、プロファイル重み調整プログラム１１９の処理手順を図１１のＰＡＤ図を用いて説明する。

【００７０】プロファイル重み調整プログラム１１９は、まずステップ１１０１において、正のプロファイル１２０に格納された文字列とその重みを一覧表示する。

【００７１】次にステップ１１０２において、負のプロファイル１２１に格納された文字列とその重みを一覧表示する。

【００７２】次にステップ１１０３において、ユーザがキーボード１０１により入力した、ユーザが重みを変更したい文字列、またはいずれかのプロファイルに追加したい文字列と、その重みを取得する。ここで、正のプロファイルにある文字列に負の重みを付与しようとした場合や、負のプロファイルにある文字列に正の重みを付与しようとした場合には、ユーザへの警告を出力するようにする等しても良い。

【００７３】次にステップ１１０４において、ステップ１１０３で取得したとおりに正のプロファイル１２０または負のプロファイル１２１の内容を変更する。

【００７４】以上が、プロファイル重み調整プログラム１１９の処理手順である。

【００７５】ここで、図１２にプロファイル重み調整プログラム１１９により、ユーザがプロファイルを調整する際にディスプレイ１００に表示する入力画面の例を示す。正のプロファイル１２０の内容が１２０１に、負のプロファイル１２１の内容が１２０２に表示される。それぞれスクロールバー１２０３および１２０４を用いて、全ての内容を表示させることも可能である。ユーザがテキストボックス１２０５に重みを変更したい文字列、またはいずれかのプロファイルに追加したい文字列を入力し、重みを１２０６に入力して送信ボタン１２０７を押下する。ここで、重みを変更したい文字列文字列はテキストボックス１２０５に入力する形ではなく、表示される一覧の中からラジオボタン等により選択する形にしても良い。

【００７６】次に、図１０に示したステップ１００１で検索条件修正制御プログラム１１６により起動される、プロファイル更新プログラム１１７の処理手順を図１３のＰＡＤ図を用いて説明する。

【００７７】プロファイル更新プログラム１１７は、まずステップ１３０１において、ユーザがキーボード１０１により入力した文書番号と、その文書番号の文書に対するユーザの評価（「所望のものであった」あるいは「所望のものでなかった」等の評価）を読み込む。

【００７８】次にステップ１３０２において、ステップ１３０１で読み込んだ文書番号に該当する文書を、テキスト１０３から登録文書保存エリア１２４に読み込む。

【００７９】次にステップ１３０３において、登録文書保存エリア１２４に格納された文書から特徴文字列を抽

出し、該文書内出現回数を計数出現頻度ファイル１０４を参照することにより抽出し、共に特徴文字列保存エリア１２５に格納する。ここで、特徴文字列の抽出方法としては前掲の「特開平１１－１４３９０２号広報」による方法を用いても良いし、形態素解析やニューラルネットワークによる学習データなどを用いる方法でもかまわない。

【００８０】次にステップ１３０４において、ステップ１３０１で読み込んだユーザの評価が正の評価であった場合には、ステップ１３０６において、特徴文字列保存エリア１２５内の文字列の出現回数を正のプロファイルの該当文字列の重みに加算する。このとき、正のプロファイル１２０に無い文字列の場合には、ステップ１３０３で読み込んだ出現回数を重みとして付与し、該文字列を正のプロファイル１２０に追加する。

【００８１】次にステップ１３０５において、ステップ１３０１で読み込んだユーザの評価が負の評価であった場合には、ステップ１３０７において、特徴文字列保存エリア１２５内の文字列の出現回数を負のプロファイルの該当文字列の重みから減算する。このとき、負のプロファイル１２１に無い文字列の場合には、ステップ１３０３で読み込んだ出現回数の負値を重みとして付与し、該文字列を負のプロファイル１２１に追加する。

【００８２】ここでステップ１３０６、１３０７において重みの加減算の方法は、ユーザの評価により調整しても良い。例えばステップ１３０６において、ユーザが「所望のものである」という評価をした場合には、その文書内の特徴文字列の出現回数を、そのまま正のプロファイル１２０の該文字列の重みに足し、「やや所望のものである」という評価をした場合には、その文書内の特徴文字列の出現回数の半数を、正のプロファイル１２０の該文字列の重みに足す、などといった方法にしても良い。また、ステップ１３０６およびステップ１３０７で重みを加減算する特徴文字列は、ステップ１３０３において抽出した出現回数の上位所定数に限定しても構わない。

【００８３】以上が、プロファイル更新プログラム１１７の処理手順である。

【００８４】次に、図１０に示したステップ１００２において検索条件修正制御プログラム１１６により起動される、検索使用文字列選択プログラム１１８の処理手順を図１４のＰＡＤ図を用いて説明する。

【００８５】検索使用文字列選択プログラム１１８は、まずステップ１４０１において、総合プロファイル１２２の内容をクリアする。

【００８６】次にステップ１４０２において、正のプロファイル１２０の中の特徴文字列のうち重みの上位所定件を抽出し、その重みと共に総合プロファイル１２２に追加する。

【００８７】次にステップ１４０３において、負のプロファイル１２１の中の特徴文字列のうち、重みの絶対値の上位所定件のもので、かつ正のプロファイル１２０の中の特徴文字列の重みの上位所定件に含まれないものを、総合プロファイル１２２に追加する。

【００８８】ここでステップ１４０２、ステップ１４０３で使用する所定件数はそれぞれ異なった値でも良い。

【００８９】以上が検索使用文字列選択プログラム１１８の処理手順である。

【００９０】以上が、本実施例における各プログラムの処理手順である。

【００９１】以下、本実施例において検索結果文書に対しユーザが負の評価をした場合の、検索条件の修正および再検索処理の流れを、図１５を用いて説明する。

【００９２】本図においては、ユーザが「高校野球」に関する文書を検索したいものとし、最初に種文書に指定した「サッカーに続き、高校野球が開幕した…」という文書１５０１から抽出された「サッカー」「高校」「野球」「開幕」という文字列１５０２が検索条件生成プログラム１１３により、正のプロファイル１２０に登録されているものとする。

【００９３】ここで、「高校サッカーが開幕した･･･」という検索結果文書１５０３に対して負の評価をした場合を想定する。

【００９４】まず、出現頻度ファイル１０４に格納された出現頻度情報のうち、ユーザが負の評価をした「高校サッカーが開幕した･･･」という文書１５０３から特徴文字列１５０４を抽出し、それぞれの特徴文字列の文書１５０３内の出現頻度とともに特徴文字列保存エリア１２５に読み込む。本図の例では、「高校」、「サッカー」、「開幕」、･･･という文字列とその出現頻度を読み込む。

【００９５】次に、特徴文字列保存エリア１２５の文字列のうち負のプロファイル１２１にある文字列についてはその重みを減算し、負のプロファイル１２１に無い文字列については、その出現回数の負の数を重みとして負のプロファイル１２１に登録する。本図の例では、「高校」、「サッカー」、「開幕」、…という文字列にそれぞれ重み「－４」、「－４」、「－１」、…を付与して負のプロファイル１２１に追加する。

【００９６】次に、正にプロファイル１２０の文字列のうち重みの上位所定数もの１５０５と、負のプロファイル１２１のうち重みの下位所定数１５０６に含まれ、かつ正のプロファイル１２０の文字列のうち上位所定数のもの１５０７に含まれないものを、総合プロファイル１２２に登録する。本図に示した例では、正のプロファイル１２０から「高校」と「野球」、負のプロファイル１２１から「サッカー」という文字列を選択し、総合プロファイル１２２に追加する。

【００９７】検索時には、この総合プロファイル１２２の文字列とその重みにより検索を行なう。本図に示した

例では、負のプロファイル中の「高校」という文字列に関する重み値－４は検索に使用されないことになる。このことにより、「高校サッカー」の文書に負の評価をしても、「高校」という文字列の重みが下がらないため、「高校野球」よりも「プロ野球」の文書に高い類似度が算出されてしまうといった問題を防ぐことができる。

【００９８】以上が、検索結果文書に対しユーザが負の評価をした場合の、検索条件の修正および再検索処理の流れである。

【００９９】以上示したように本実施例によれば、ユーザが「所望のものでない」と評価した文書から抽出された文字列のうち、ユーザが「所望のものである」と評価した文書から抽出された文字列を、重みを下げる対象から除外する形態をとる。そのため、ユーザの所望ではない概念を表す文字列のみの重みを適切に減算することができる。したがって、ユーザが「所望のものでない」と評価した文書から抽出した文字列の重みを単純に減算すると、ユーザの所望の概念を表す文字列の重みまで減算してしまい、検索結果が改善しない、といった問題を解決できる。

【０１００】また、本実施例によれば、検索結果文書の内容を表示する際、検索条件データに保存されている文字列の重み正負により文字列を別の形式でハイライト表示する形態をとる。

【０１０１】この方法により、ユーザは、検索結果文書がどの程度所望の内容を示しているかを視覚的に容易に判断できる。また、正の重みが付与された文字列や負の重みが付与された文字列として、どのようなものが所望文書やノイズ文書に含まれているかを見ることにより、次回以降のプロファイルの調整に役立てることができるようになる。

【０１０２】また、本実施例によれば、検索条件データの中の文字列のうち検索に用いる文字列をユーザが選択、あるいはそれぞれの文字列の重みをユーザが調整する形態をとる。

【０１０３】この方法により、ユーザの所望する内容を特徴付けるものでないものを、検索に使用することを防ぐことができ、適切な検索結果を得られるようになる。

【０１０４】図１３に示したプロファイル更新プログラムの処理おいては、ユーザが負の評価をした際に、評価対象文書から抽出した文字列を負のプロファイル１２１に追加した後、総合プロファイル１２２に追加する文字列を選択する形態をとっている。ここで図１６に示すように、評価対象文書から抽出した文字列のうち、負のプロファイル１２１に追加する文字列を選択する形態をとっても良い。

【０１０５】すなわち、図１６のステップ１３０５において、ステップ１３０１で読み込んだユーザの評価が負の評価であった場合には、ステップ１３０７を実行する前に図１６に示すプロファイル更新用文字列選択ステップ１６０１を実行しても良い。ここでプロファイル更新用文字列選択ステップ１６０１は、特徴文字列保存エリア１２５の文字列のうち、正のプロファイル１２０中の重みの上位のものに含まれるものを、特徴文字列保存エリア１２５からクリアするステップである。これにより、正のプロファイル１２０に追加されているユーザの所望の概念を表す文字列に、負の重みを付与し負のプロファイル１２１に追加してしまうことを防ぐことができる。

【０１０６】以下、本発明の第二の実施例について説明する。

【０１０７】第一の実施例においては、検索時に使用する文字列、または検索条件の修正時にプロファイルに追加する文字列をシステムが自動的に選択する。したがって、検索結果文書に対するユーザの評価が不適切な場合には、検索精度が向上しないという問題がある。

【０１０８】以上の問題を解決するために、本発明の第二の実施例では、ユーザが正または負の評価をした文書から抽出される文字列を一覧表示し、正の重みまたは負の重みを付与する文字列をユーザが選択する手段を提供するものである。

【０１０９】本実施例は図１に示す第一の実施例とほぼ同様の構成をとる。ここで図１７に示すように検索条件修正制御プログラム１１６ａはプロファイル更新用文字列ユーザ選択プログラム１７０１、プロファイル更新プログラム１１７ａ、および検索使用文字列選択プログラム１１８により構成される。また、図１８に示すようにプロファイル更新プログラム１１７ａの処理手順が、第一の実施例におけるプロファイル更新プログラム１１７と異なる。

【０１１０】以下、第二の実施例における、プロファイル更新プログラム１１７ａの処理手順について図１８のＰＡＤ図を用いて説明する。

【０１１１】まずプロファイル更新プログラム１１７ａは、まずステップ１８０１において、ユーザがキーボード１０１により入力した文書番号と、その文書番号の文書に対するユーザの評価（「所望のものであった」あるいは「所望のものでなかった」等の評価）を読み込む。

【０１１２】次にステップ１８０２において、ステップ１８０１で読み込んだ文書番号に該当する文書を、テキスト１０３から登録文書保存エリア１２４に読み込む。

【０１１３】次にステップ１８０３において、登録文書保存エリア１２４に格納された文書から特徴文字列を抽出し、該文書内出現回数を計数出現頻度ファイル１０４を参照することにより抽出し、共に特徴文字列保存エリア１２５に格納する。ここで、特徴文字列の抽出方法としては前掲の「特開平１１－１４３９０２号広報」による方法を用いても良いし、形態素解析やニューラルネットワークによる学習データなどを用いる方法でもかまわない。

【０１１４】次にステップ１８０４において、プロファイル更新用文字列ユーザ選択プログラム１７０１を起動し、ステップ１８０３において読み込んだ文字列のうちユーザが選択しなかった文字列を、特徴文字列保存エリア１２５からクリアする。

【０１１５】次にステップ１８０５において、ステップ１８０１で読み込んだユーザの評価が正の評価であった場合には、ステップ１８０７において、特徴文字列保存エリア１２５の文字列の出現回数を正のプロファイルの該当文字列の重みに加算する。このとき、正のプロファイル１２０に無い文字列の場合には、ステップ１８０３で読み込んだ出現回数を重みとして付与し、該文字列を正のプロファイル１２０に追加する。

【０１１６】次にステップ１８０６において、ステップ１８０１で読み込んだユーザの評価が負の評価であった場合には、ステップ１８０８において、特徴文字列保存エリア１２５の文字列の出現回数を負のプロファイルの該当文字列の重みから減算する。このとき、負のプロファイル１２１に無い文字列の場合には、ステップ１８０３で読み込んだ出現回数の負値を重みとして付与し、該文字列を負のプロファイル１２１に追加する。

【０１１７】ここでステップ１８０７、１８０８において重みの加減算の方法は、ユーザの評価により調整しても良い。例えばステップ１８０７において、ユーザが「所望のものである」という評価をした場合には、その文書内の特徴文字列の出現回数を、そのまま正のプロファイル１２０の該文字列の重みに足し、「やや所望のものである」という評価をした場合には、その文書内の特徴文字列の出現回数の半数を、正のプロファイル１２０の該文字列の重みに足す、などといった方法にしても良い。また、ステップ１８０７およびステップ１８０８で重みを加減算する特徴文字列は、ステップ１８０３において抽出した出現回数の上位所定数に限定しても構わない。

【０１１８】以上が、プロファイル更新プログラム１１７ａの処理手順である。

【０１１９】次に図１８に示したステップ１８０４でプロファイル更新プログラム１１７ａにより起動される、プロファイル更新用文字列ユーザ選択プログラム１７０１の処理手順を、図１９のＰＡＤ図を用いて説明する。

【０１２０】まずステップ１９０１において、特徴文字列保存エリア１２５内の特徴文字列を一覧表示する。

【０１２１】次にステップ１９０２において、ステップ１９０１で表示した文字列のうち、ユーザが選択しなかった文字列を取得し、該文字列の情報を特徴文字列保存エリア１２５からクリアする。

【０１２２】以上がプロファイル更新用文字列ユーザ選択プログラム１７０１の処理手順である。

【０１２３】ここで、プロファイル更新用文字列ユーザ選択プログラム１７０１により、ユーザがプロファイルに追加したい文字列を選択する画面の例を図２０に示す。ウィンドウ２００１に、ユーザが評価した文書から抽出される特徴文字列がチェックボックスと共に表示される。特徴文字列が多数ある場合はスクロールバー２００２を用いてすべての文字列をウィンドウ２００１内で参照することができる。ユーザは、ウィンドウ２００１内の文字列のうち、プロファイルに追加したい文字列のチェックボックスをチェックし、送信ボタン２００３を押下する。

【０１２４】なお、文字列の選択方法は図２０の例のようにチェックボックスを用いたものでも良いし、各文字列に識別番号を付与して識別番号と共に一覧表示するようにし、文字列の識別番号により選択する方法でも良い。

【０１２５】以下、本実施例において検索結果テキストに対しユーザが負の評価をした場合の、検索条件の修正および再検索処理の流れを、図２１を用いて説明する。

【０１２６】本図においては、ユーザが「高校野球」に関するテキストを検索したいものとし、最初に種文書に指定した「サッカーに続き、高校野球が開幕した…」というテキスト２１０１から抽出されたサッカー」「高校」「野球」「開幕」という文字列２１０２が検索条件生成プログラム１１３により、正のプロファイル１２０に登録されているものとする。

【０１２７】ここで、「高校サッカーの１回戦が･･･」という検索結果テキストに対して負の評価をした場合を想定する。

【０１２８】まず、出現頻度ファイル１０４に格納された出現頻度情報のうち、ユーザが負の評価をした「高校サッカーの１回戦が･･･」という文書２１０３から特徴文字列２１０４を抽出し、それぞれの特徴文字列の文書２１０３内の出現頻度とともに特徴文字列保存エリア１２５に読み込む。本図の例では、「高校」、「サッカー」、「１回戦」、･･･という文字列とその出現頻度が読み込まれる。

【０１２９】次に、前述した図２０の画面でユーザが選択した文字列の情報を、文字列保存エリア１２５からクリアする。本図の例では、ユーザが「高校野球」に関するテキストを所望しており、「サッカー」に関するテキストは所望ではない。したがってユーザは「サッカー」という文字列のみに負の重みを加えると指定するものとする。このとき、文字列保存エリア１２５から、「高校」および「１回戦」という文字列とその重みをクリアする。

【０１３０】次に、出現頻度情報２１０４のうち負のプロファイル１２１にある文字列についてはその重みを減算し、負のプロファイル１２１に無い文字列については、その出現回数の負の数を重みとして負のプロファイル１２１に登録する。本図の例では、「サッカー」という文字列に重み「－４」を付与して正のプロファイル１

２０に追加する。

【０１３１】次に、正にプロファイル１２０の文字列のうち重みの上位所定数もの２１０５と、負のプロファイル１２１のうち重みの下位所定数２１０６に含まれ、かつ正のプロファイル１２０の文字列のうち上位所定数のもの２１０７に含まれないものを、総合プロファイル１２２に登録する。検索時には、この総合プロファイル１２２の文字列とその重みにより検索を行なう。

【０１３２】以上のように、本図に示した例では、「高校サッカーの１回戦が…」というテキストに負の評価をしても、「高校」という文字列の重みが下がらないため、「高校野球」よりも「プロ野球」のテキストに高い類似度が算出されてしまうといった問題を防ぐことができる。また、正のプロファイル１２０に無い「１回戦」という文字列の重みがさがらないため、「高校野球の１回戦」といったユーザが所望するテキストの類似度が下がってしまうといった問題を防ぐことができる。

【０１３３】以上が、検索結果テキストに対しユーザが負の評価をした場合の、検索条件の修正および再検索処理の流れである。

【０１３４】なお、本実施例において検索結果文書に対しユーザが正の評価をした場合にも同様に、正のプロファイルに追加する文字列を選択することができる。したがって、正の評価をした文書から抽出されるがユーザの概念を表す文字列ではない文字列に、正の重みを付与してしまうことを防ぐことができる。

【０１３５】以上が、本発明の第二の実施例である。

【０１３６】以上示したように本実施例によれば、ユーザが「所望のものでない」と評価した文書から抽出された文字列のうち、ユーザが所望する概念を表す文字列をユーザが指定することにより、該文字列を重みを下げる対象から除外する形態をとる。そのため、ユーザの所望ではない概念を表す文字列のみの重みを適切に減算することができる。したがって、ユーザが「所望のものでない」と評価した文書から抽出した文字列の重みを単純に減算すると、ユーザの所望の概念を表す文字列の重みまで減算してしまい、検索結果が改善しない、といった問題を解決できる。

【０１３７】また、ユーザが「所望のものである」と評価した文書から抽出された文字列のうち、ユーザが所望する概念を表さない文字列をユーザが指定することにより、該文字列を重みを上げる対象から除外する形態をとる。そのため、ユーザの所望する概念を表す文字列のみの重みを適切に加算することができる。したがって、ユーザが「所望のものである」と評価した文書から抽出した文字列の重みを単純に加算すると、ユーザの所望の概念を表さない文字列の重みまで加算してしまい、検索結果が改善しない、といった問題を解決できる。

【０１３８】なお、第一、第二の実施例において、ひとつの検索結果文書に対しユーザが評価を入力し、その評価を反映した検索結果を出力するようにしたが、複数の検索結果文書に対しそれぞれ異なった評価を一度に入力し、それらの評価を反映した検索結果を出力するようにしても構わない。

【０１３９】また、第一、第二の実施例において、最初に種文書を設定し、その種文書に類似した内容を持つ文書を検索するものとしたが、最初にキーワードを設定する全文検索を行なう形式にしても良い。その場合には、図７に示した検索条件生成プログラム１１３のステップ７０２、７０３のかわりに、入力したキーワードを所定の重みを付与して正のプロファイル１２０、および総合プロファイル１２２に追加すれば良い。

【０１４０】本実施例によれば、ユーザの所望の概念を表す単語の重みを減算しないため、ユーザが「所望のものでない」といった評価を与えた検索結果文書から抽出した情報をもとに検索結果を改善することができる。

【０１４１】

【発明の効果】本発明によれば、ユーザが「所望のものでない」といった評価を与えた文書から抽出した情報のうち適切なものを使用して、検索結果を改善することができる。

【図面の簡単な説明】

【図１】本発明の第一の実施例の構成を示す図である。

【図２】従来技術によるレリバンスフィードバック処理の例を示す図である。

【図３】従来技術によるレリバンスフィードバック処理により検索結果が改善しない例を示す図である。

【図４】本発明の第一の実施例におけるシステム制御プログラム１１０の処理手順を示すＰＡＤ図である。

【図５】本発明の第一の実施例における文書登録プログラム１１１の処理手順を示すＰＡＤ図である。

【図６】本発明の第一の実施例における検索制御プログラム１１２の処理手順を示すＰＡＤ図である。

【図７】本発明の第一の実施例における検索条件生成プログラム１１３の処理手順を示すＰＡＤ図である。

【図８】本発明の第一の実施例における類似文書検索プログラム１１４の処理手順を示すＰＡＤ図である。

【図９】本発明の第一の実施例における検索結果文書内容表示プログラム１１５の処理手順を示すＰＡＤ図である。

【図１０】本発明の第一の実施例における検索条件修正制御プログラム１１６の処理手順を示すＰＡＤ図である。

【図１１】本発明の第一の実施例におけるプロファイル重み調整プログラム１１９の処理手順を示すＰＡＤ図である。

【図１２】本発明の第一の実施例において、ユーザがプロファイルを調整する際にディスプレイ１００に表示する入力画面の例を示す図である。

【図１３】本発明の第一の実施例におけるプロファイル

更新プログラム１１７の処理手順を示すＰＡＤ図である。

【図１４】本発明の第一の実施例における検索使用文字列選択プログラム１１８の処理手順を示すＰＡＤ図である。

【図１５】本発明の第一の実施例において、検索結果文書に対しユーザが負の評価をした場合の、検索条件の修正および再検索処理の流れを示す図である。

【図１６】本発明の第一の実施例におけるプロファイル更新プログラムの処理１１７の処理の一形態を示すＰＡＤ図である。

【図１７】本発明の第二の実施例における検索条件修正プログラム１１６ａの構成を示すＰＡＤ図である。

【図１８】本発明の第二の実施例におけるプロファイル更新プログラム１１７ａの処理手順を示すＰＡＤ図である。

【図１９】本発明の第二の実施例におけるプロファイル更新用文字列ユーザ選択プログラム１７０１の処理手順を示すＰＡＤ図である。

【図２０】本発明の第二の実施例において、ユーザがプロファイルに追加したい文字列を選択する画面の例を示すＰＡＤ図である。

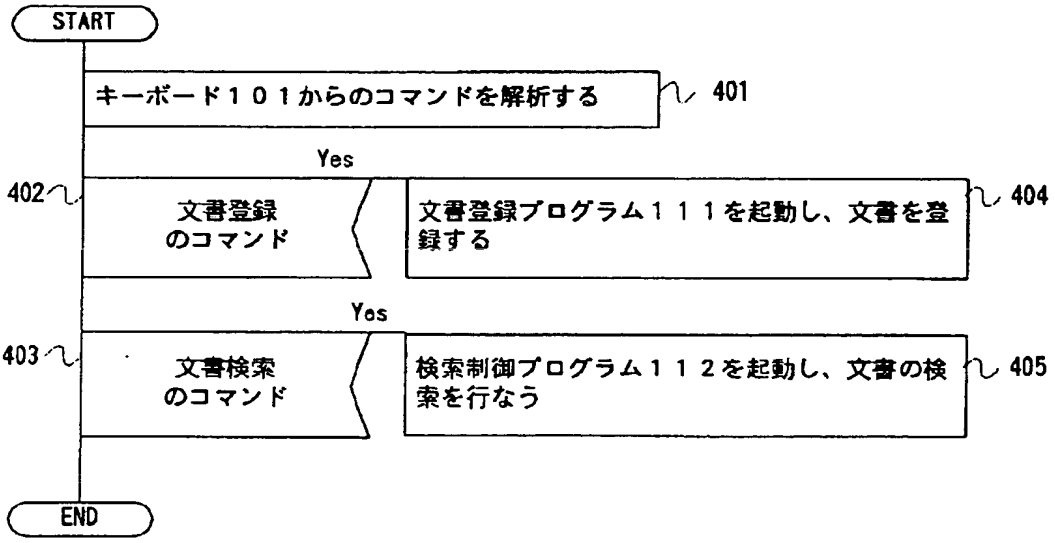
【図２１】本発明の第二の実施例において、検索結果文書に対しユーザが負の評価をした場合の、検索条件の修正および再検索処理の流れを示す図である。

【符号の説明】

１００ ディスプレイ
１０１ キーボード
１０２ 中央演算処理装置（ＣＰＵ）
１０３ テキスト
１０４ 出現頻度ファイル
１０５ 磁気ディスク装置
１０６ フロッピディスクドライブ（ＦＤＤ）
１０７ フロッピディスク
１０８ バス
１０９ 主メモリ
１１０ システム制御プログラム
１１１ 文書登録プログラム
１１２ 検索制御プログラム
１１３ 検索条件生成プログラム
１１４ 類似文書検索プログラム
１１５ 検索結果文書内容表示プログラム
１１６ 検索条件修正制御プログラム
１１７ プロファイル更新プログラム
１１８ 検索使用文字列選択プログラム
１１９ プロファイル重み調整プログラム
１２０ 正のプロファイル
１２１ 負のプロファイル
１２２ 総合プロファイル
１２３ 種文書保存エリア
１２４ 登録文書保存エリア
１２５ 特徴文字列保存エリア
１２６ 表示文書保存エリア

【図４】

図４



【図１】

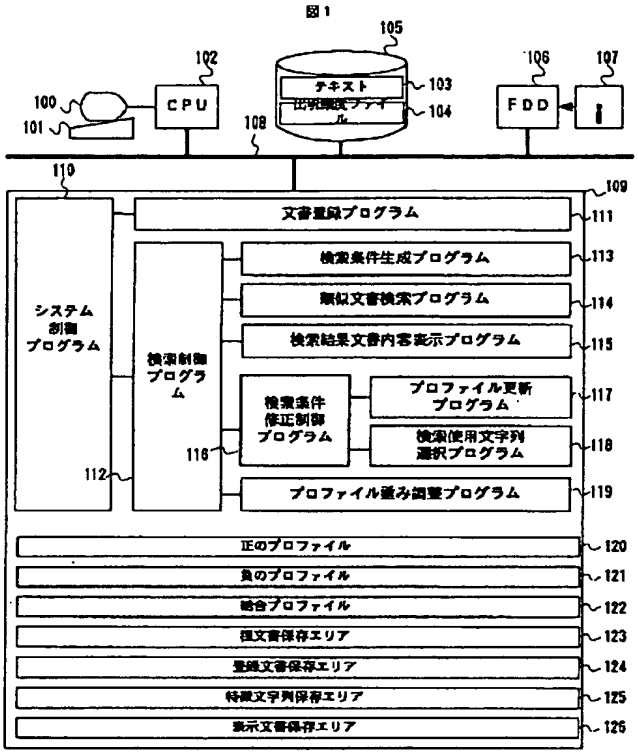


図１
105
102
100
101
CPU
テキスト
出現頻度ファイル
103
104
106
FDD
107
108
110
109
システム制御プログラム
文書登録プログラム
111
検索制御プログラム
検索条件生成プログラム
113
類似文書検索プログラム
114
検索結果文書内容表示プログラム
115
検索条件修正制御プログラム
プロファイル更新プログラム
117
検索使用文字列選択プログラム
118
116
112
プロファイル重み調整プログラム
119
正のプロファイル
120
負のプロファイル
121
統合プロファイル
122
種文書保存エリア
123
登録文書保存エリア
124
特徴文字列保存エリア
125
表示文書保存エリア
126

【図１２】

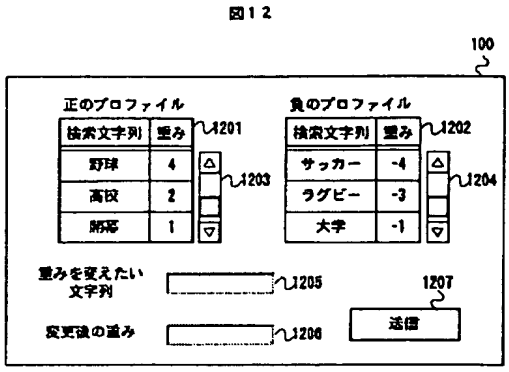
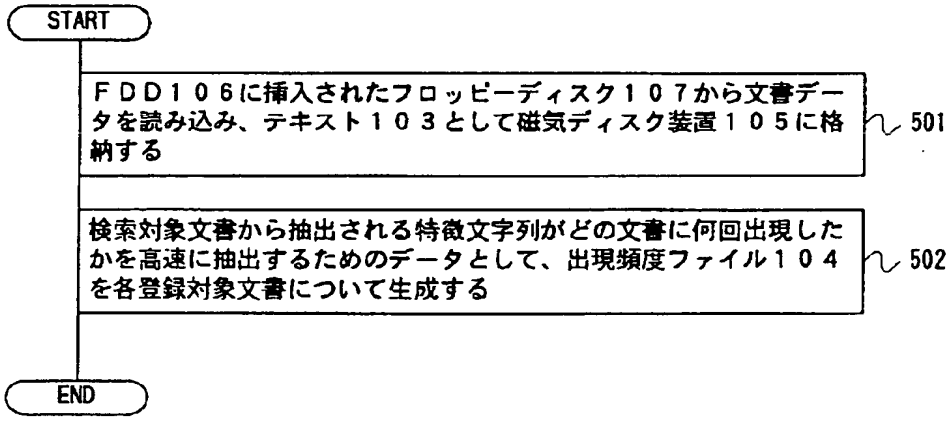


図１２
100
正のプロファイル
検索文字列
重み
1201
野球
4
高校
2
箱根
1
1203
負のプロファイル
検索文字列
重み
1202
サッカー
-4
ラグビー
-3
大学
-1
1204
重みを変えたい文字列
1205
変更後の重み
1206
1207
送信

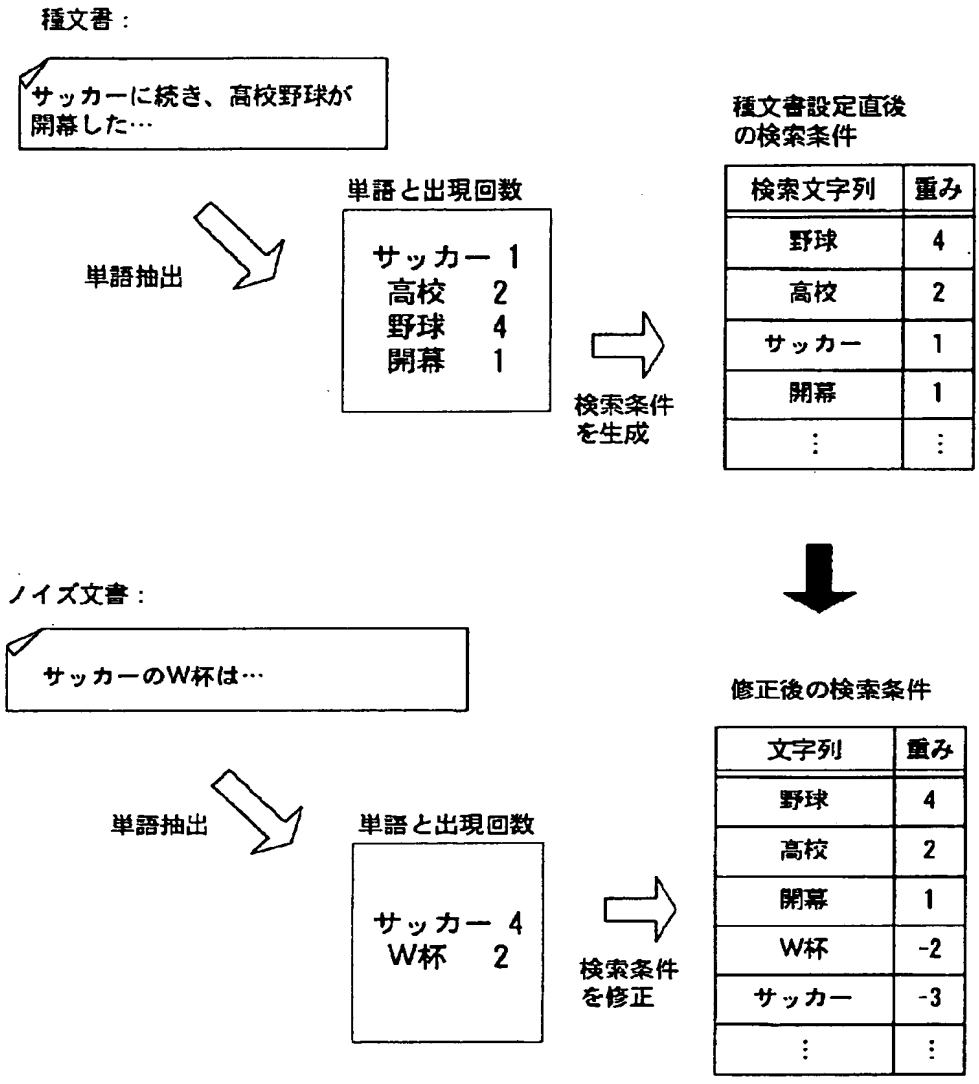
【図５】

図５


START
ＦＤＤ１０６に挿入されたフロッピーディスク１０７から文書データを読み込み、テキスト１０３として磁気ディスク装置１０５に格納する
501
検索対象文書から抽出される特徴文字列がどの文書に何回出現したかを高速に抽出するためのデータとして、出現頻度ファイル１０４を各登録対象文書について生成する
502
END

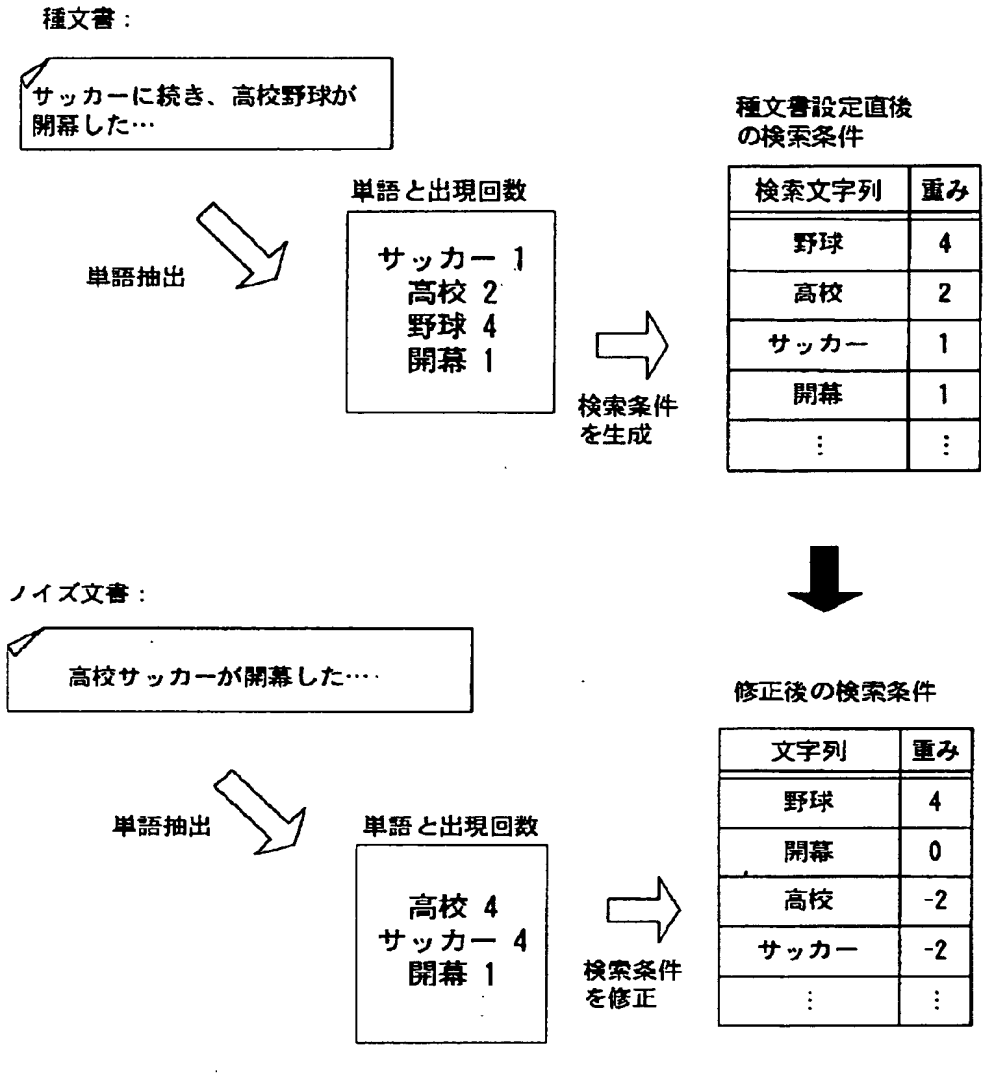
【図２】

図２


種文書：
サッカーに続き、高校野球が開幕した…
単語抽出
単語と出現回数
サッカー 1
高校 2
野球 4
開幕 1
検索条件を生成
種文書設定直後の検索条件
検索文字列
重み
野球 4
高校 2
サッカー 1
開幕 1
ノイズ文書：
サッカーのW杯は…
単語抽出
単語と出現回数
サッカー 4
W杯 2
検索条件を修正
修正後の検索条件
文字列
重み
野球 4
高校 2
開幕 1
W杯 -2
サッカー -3

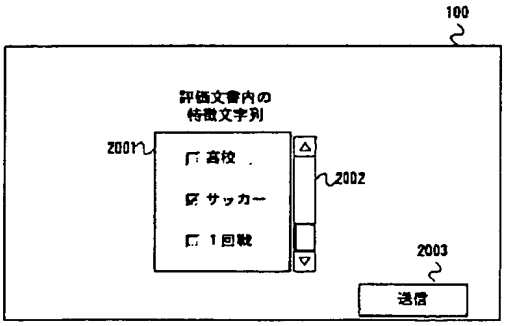
【図３】

図３


種文書：
サッカーに続き、高校野球が開幕した…
単語抽出
単語と出現回数
サッカー 1
高校 2
野球 4
開幕 1
検索条件を生成
種文書設定直後の検索条件
検索文字列 重み
野球 4
高校 2
サッカー 1
開幕 1
ノイズ文書：
高校サッカーが開幕した…
単語抽出
単語と出現回数
高校 4
サッカー 4
開幕 1
検索条件を修正
修正後の検索条件
文字列 重み
野球 4
開幕 0
高校 -2
サッカー -2

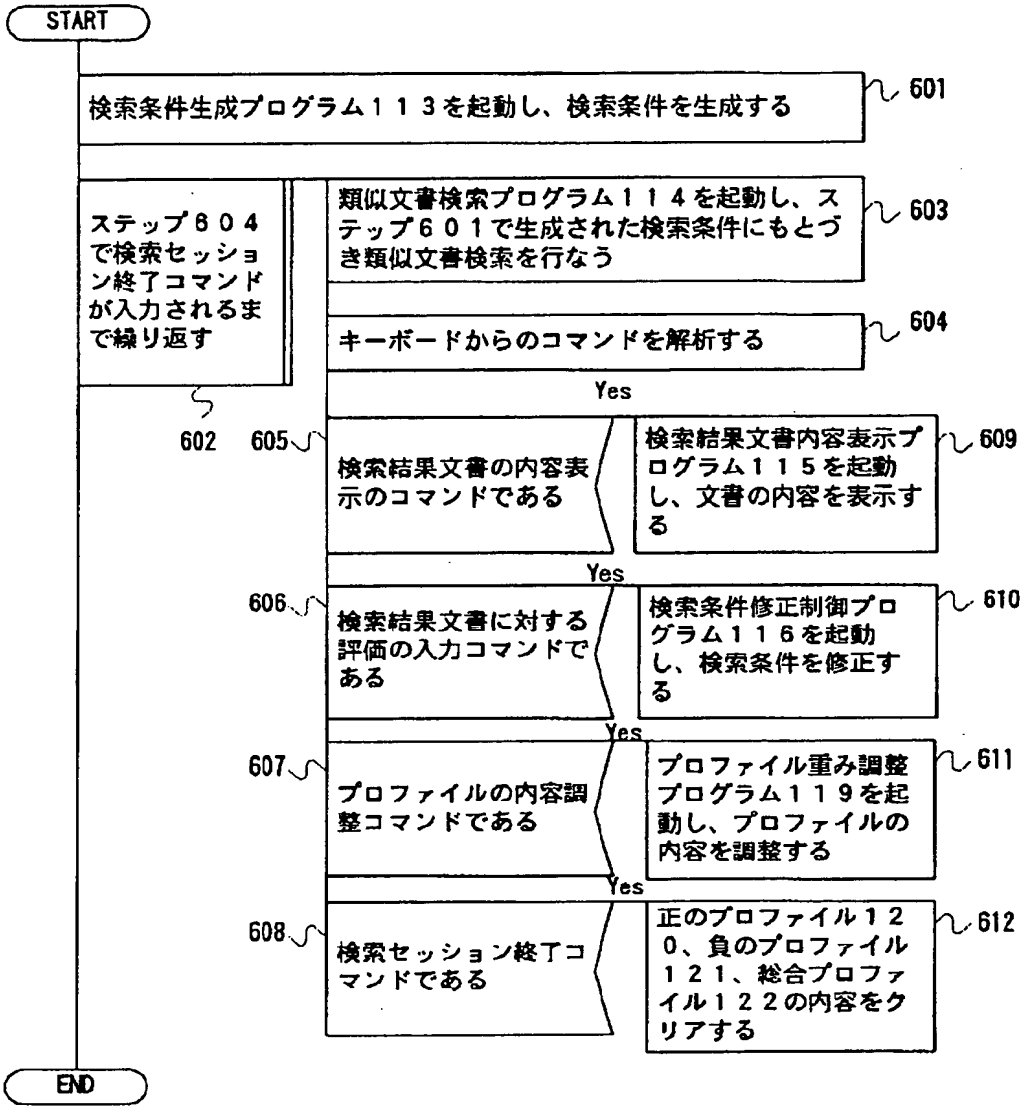
【図２０】

図２０


100
評価文書内の特徴文字列
2001
高校
サッカー
1回戦
2002
2003
送信

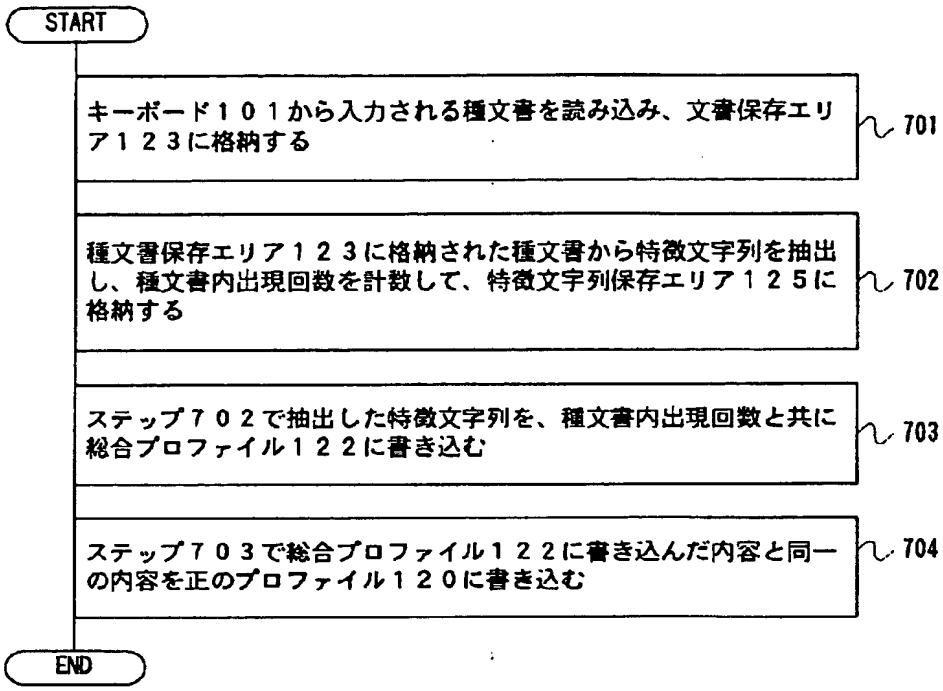
【図６】

図6


START
検索条件生成プログラム１１３を起動し、検索条件を生成する
601
ステップ６０４で検索セッション終了コマンドが入力されるまで繰り返す
602
類似文書検索プログラム１１４を起動し、ステップ６０１で生成された検索条件にもとづき類似文書検索を行なう
603
キーボードからのコマンドを解析する
604
Yes
605
検索結果文書の内容表示のコマンドである
検索結果文書内容表示プログラム１１５を起動し、文書の内容を表示する
609
Yes
606
検索結果文書に対する評価の入力コマンドである
検索条件修正制御プログラム１１６を起動し、検索条件を修正する
610
Yes
607
プロファイルの内容調整コマンドである
プロファイル重み調整プログラム１１９を起動し、プロファイルの内容を調整する
611
Yes
608
検索セッション終了コマンドである
正のプロファイル１２０、負のプロファイル１２１、総合プロファイル１２２の内容をクリアする
612
END

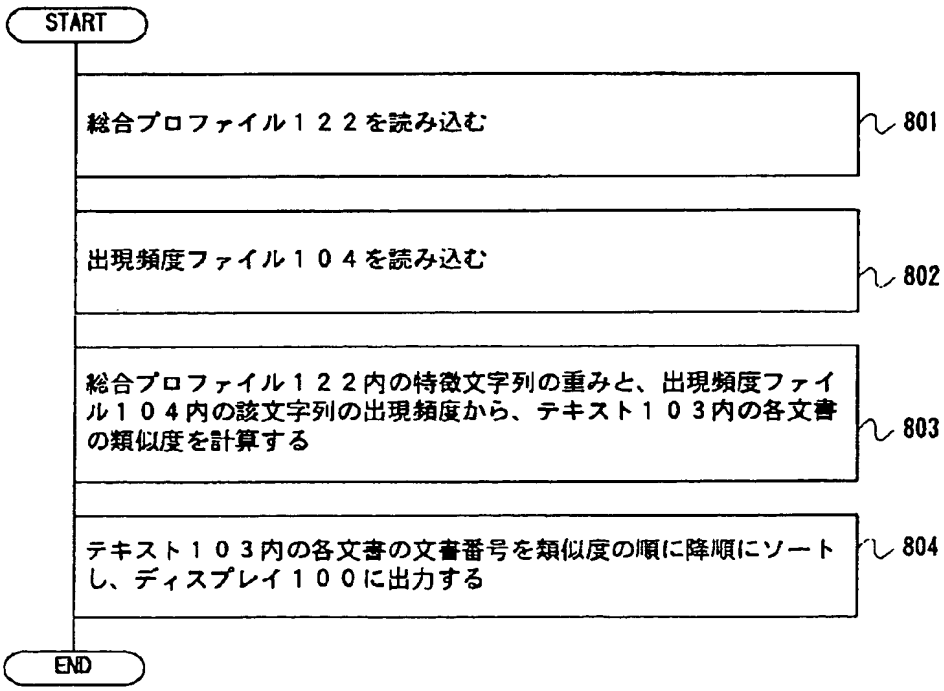
【図７】

図７


START
キーボード１０１から入力される種文書を読み込み、文書保存エリア１２３に格納する
701
種文書保存エリア１２３に格納された種文書から特徴文字列を抽出し、種文書内出現回数を計数して、特徴文字列保存エリア１２５に格納する
702
ステップ７０２で抽出した特徴文字列を、種文書内出現回数と共に総合プロファイル１２２に書き込む
703
ステップ７０３で総合プロファイル１２２に書き込んだ内容と同一の内容を正のプロファイル１２０に書き込む
704
END

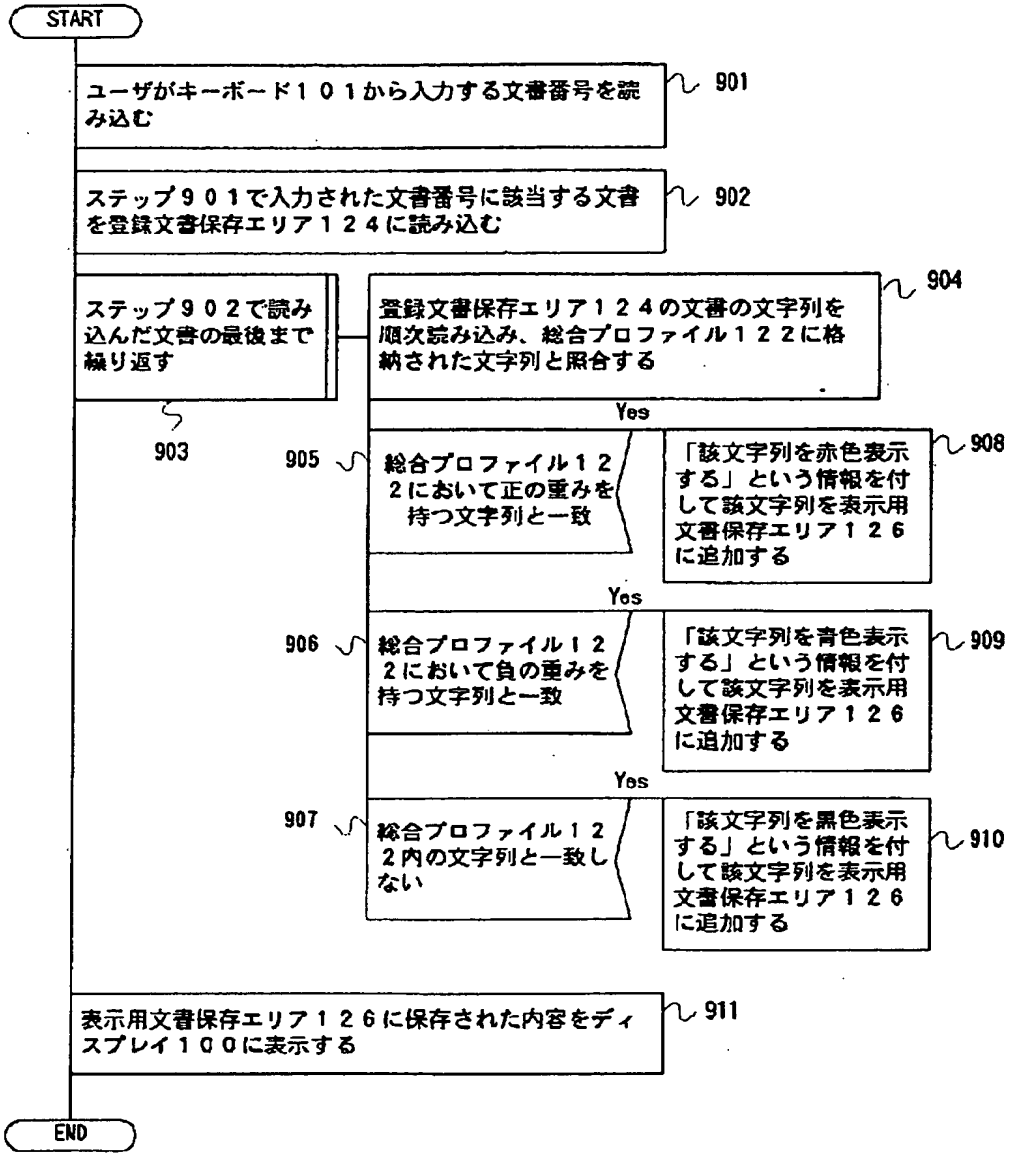
【図８】

図８


START
総合プロファイル１２２を読み込む
801
出現頻度ファイル１０４を読み込む
802
総合プロファイル１２２内の特徴文字列の重みと、出現頻度ファイル１０４内の該文字列の出現頻度から、テキスト１０３内の各文書の類似度を計算する
803
テキスト１０３内の各文書の文書番号を類似度の順に降順にソートし、ディスプレイ１００に出力する
804
END

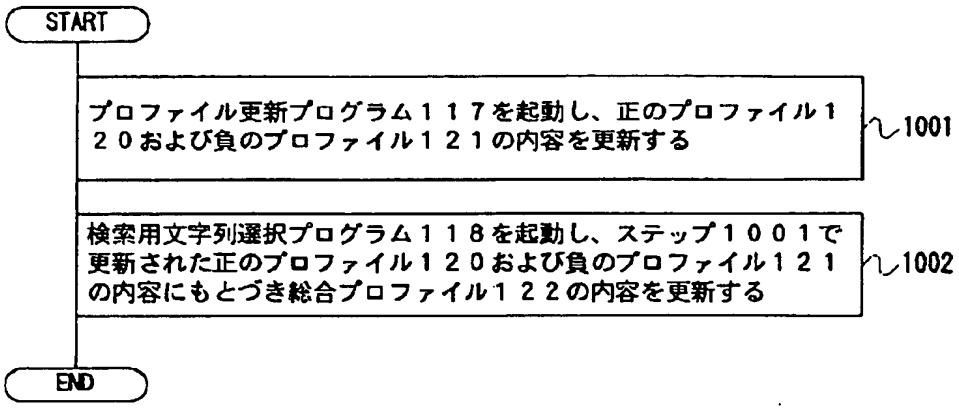
【図９】

図９


START
ユーザがキーボード１０１から入力する文書番号を読み込む
901
ステップ９０１で入力された文書番号に該当する文書を登録文書保存エリア１２４に読み込む
902
ステップ９０２で読み込んだ文書の最後まで繰り返す
903
登録文書保存エリア１２４の文書の文字列を順次読み込み、総合プロファイル１２２に格納された文字列と照合する
904
Yes
905
総合プロファイル１２２において正の重みを持つ文字列と一致
「該文字列を赤色表示する」という情報を付して該文字列を表示用文書保存エリア１２６に追加する
908
Yes
906
総合プロファイル１２２において負の重みを持つ文字列と一致
「該文字列を青色表示する」という情報を付して該文字列を表示用文書保存エリア１２６に追加する
909
Yes
907
総合プロファイル１２２内の文字列と一致しない
「該文字列を黒色表示する」という情報を付して該文字列を表示用文書保存エリア１２６に追加する
910
表示用文書保存エリア１２６に保存された内容をディスプレイ１００に表示する
911
END

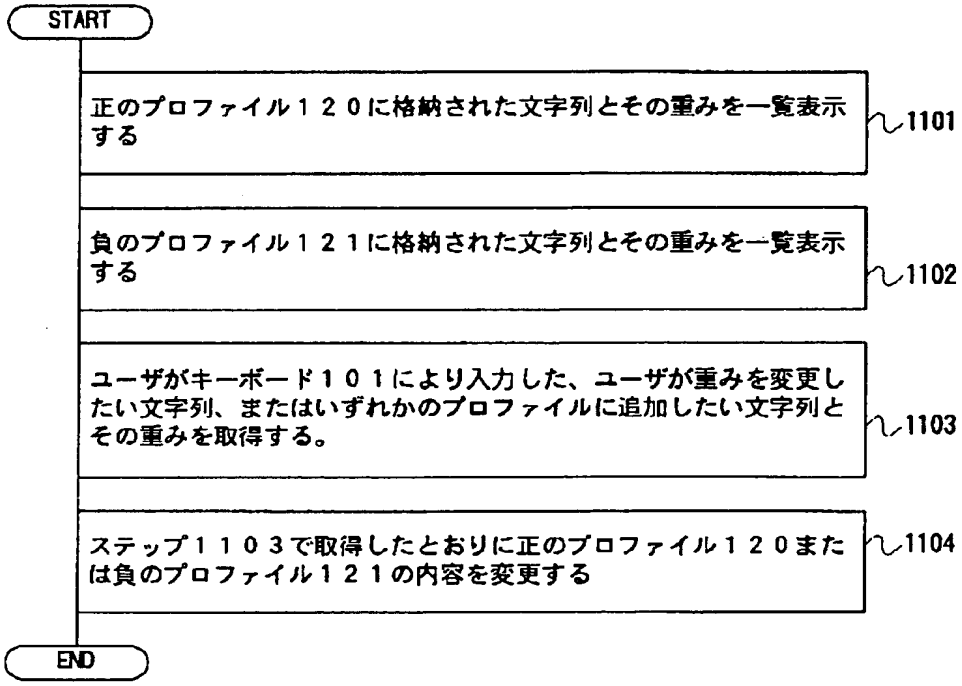
【図１０】

図１０



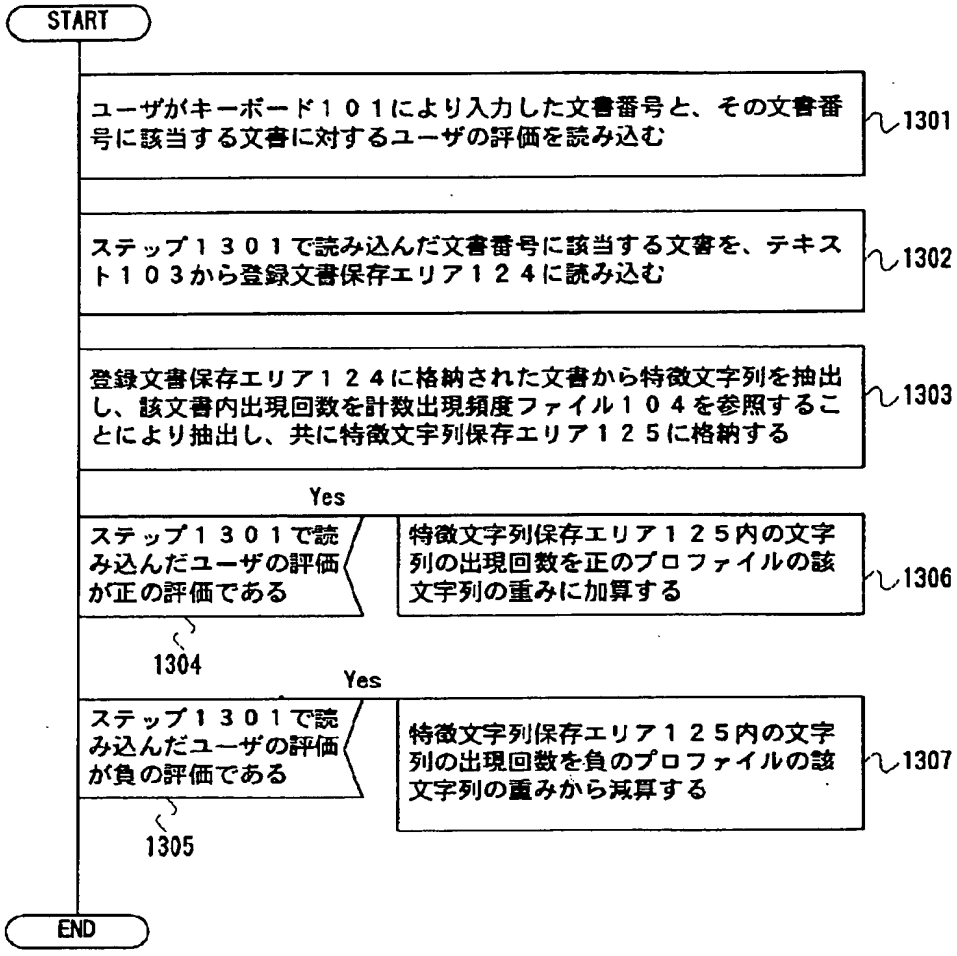
【図１１】

図１１



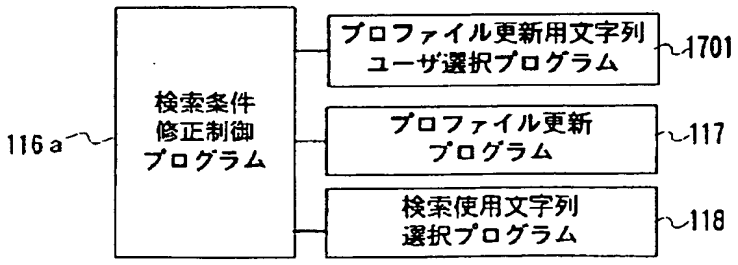
【図１３】

図１３



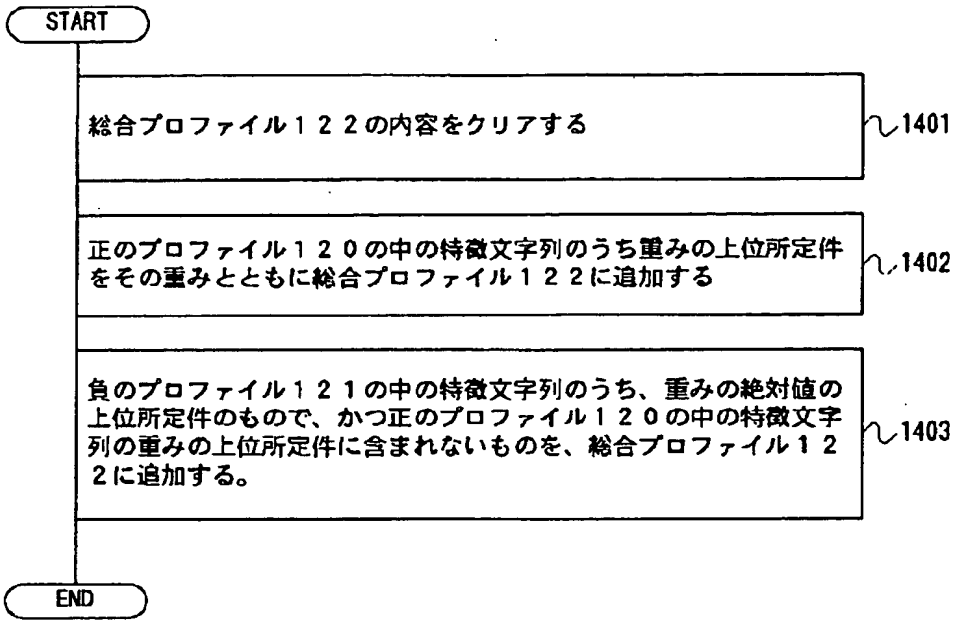
【図１７】

図１７



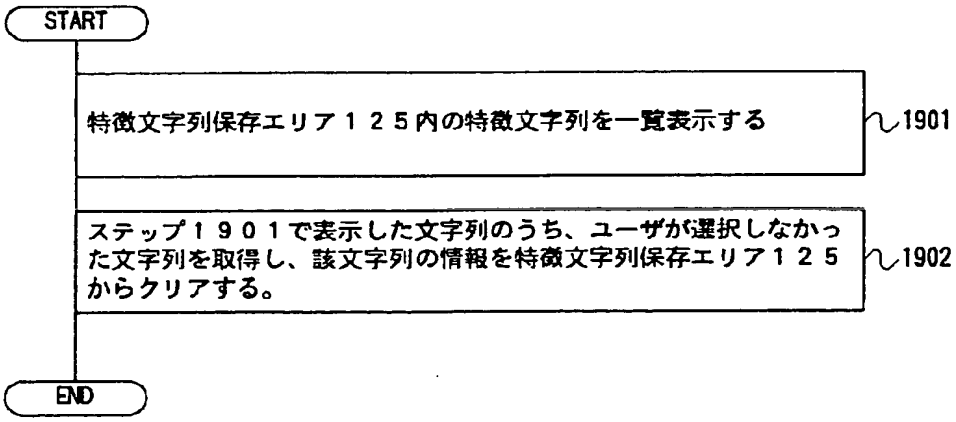
【図１４】

図１４



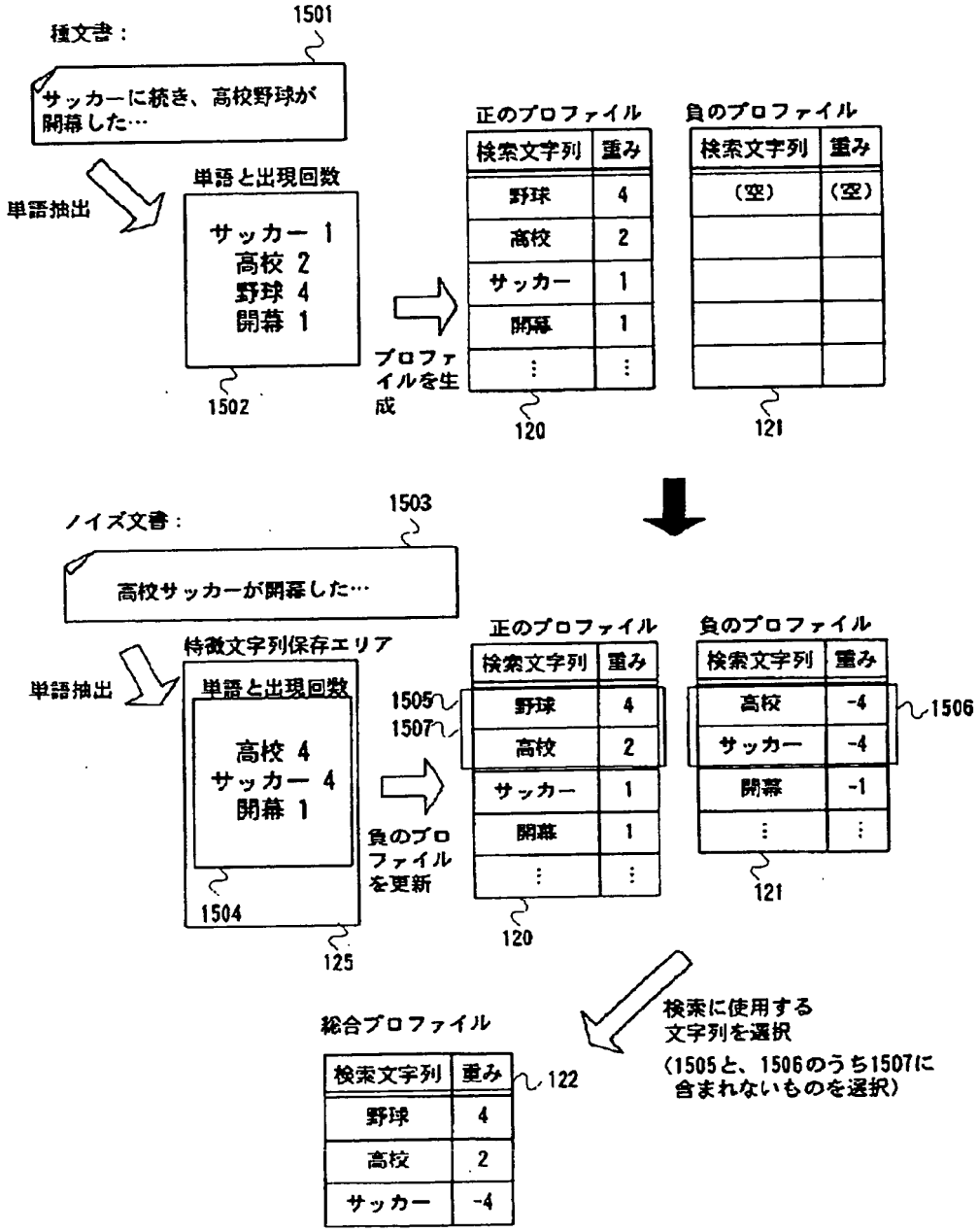
【図１９】

図１９



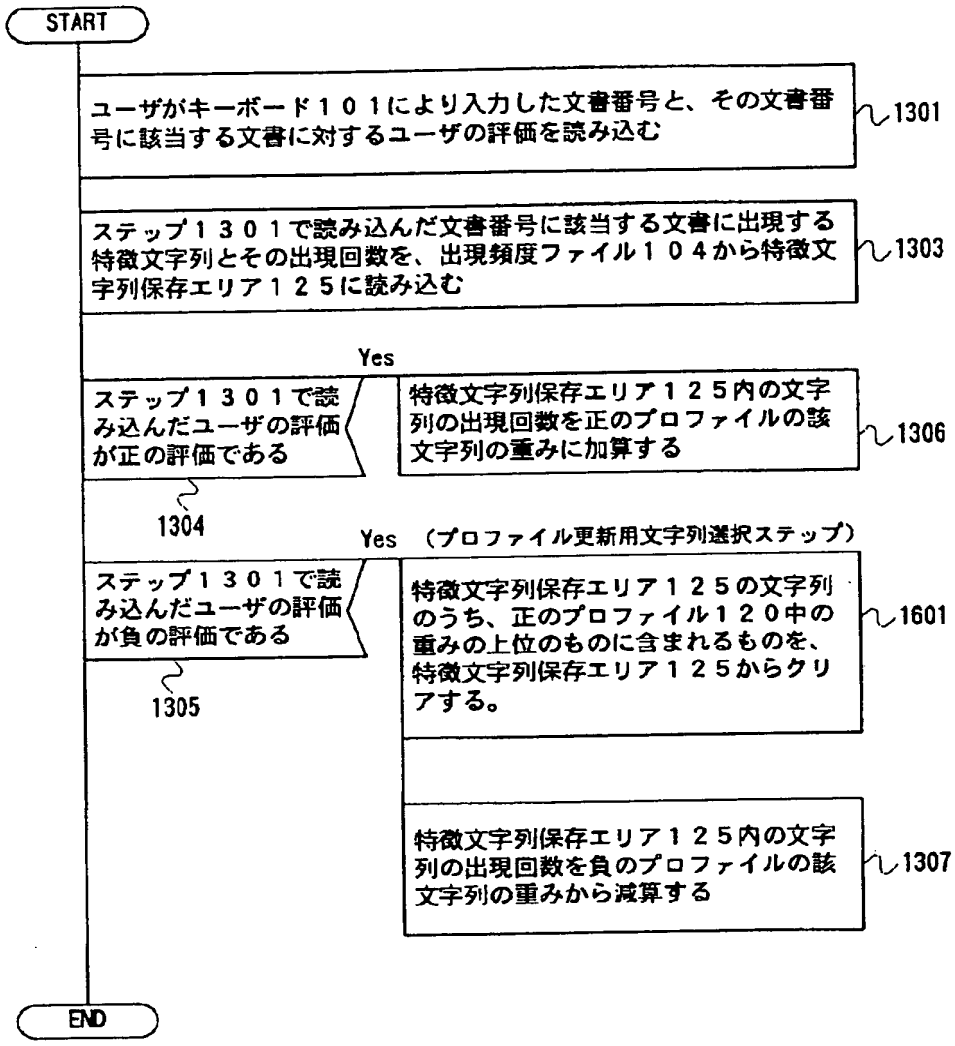
【図１５】

図１５


種文書：
1501
サッカーに続き、高校野球が開幕した…
単語抽出
単語と出現回数
サッカー 1
高校 2
野球 4
開幕 1
1502
プロファイルを生成
正のプロファイル
検索文字列 重み
野球 4
高校 2
サッカー 1
開幕 1
120
負のプロファイル
検索文字列 重み
(空) (空)
121
ノイズ文書：
1503
高校サッカーが開幕した…
単語抽出
特徴文字列保存エリア
単語と出現回数
高校 4
サッカー 4
開幕 1
1504
125
負のプロファイルを更新
正のプロファイル
検索文字列 重み
1505
1507
野球 4
高校 2
サッカー 1
開幕 1
120
負のプロファイル
検索文字列 重み
高校 -4
サッカー -4
開幕 -1
1506
121
総合プロファイル
検索文字列 重み
野球 4
高校 2
サッカー -4
122
検索に使用する文字列を選択
(1505と、1506のうち1507に含まれないものを選択)

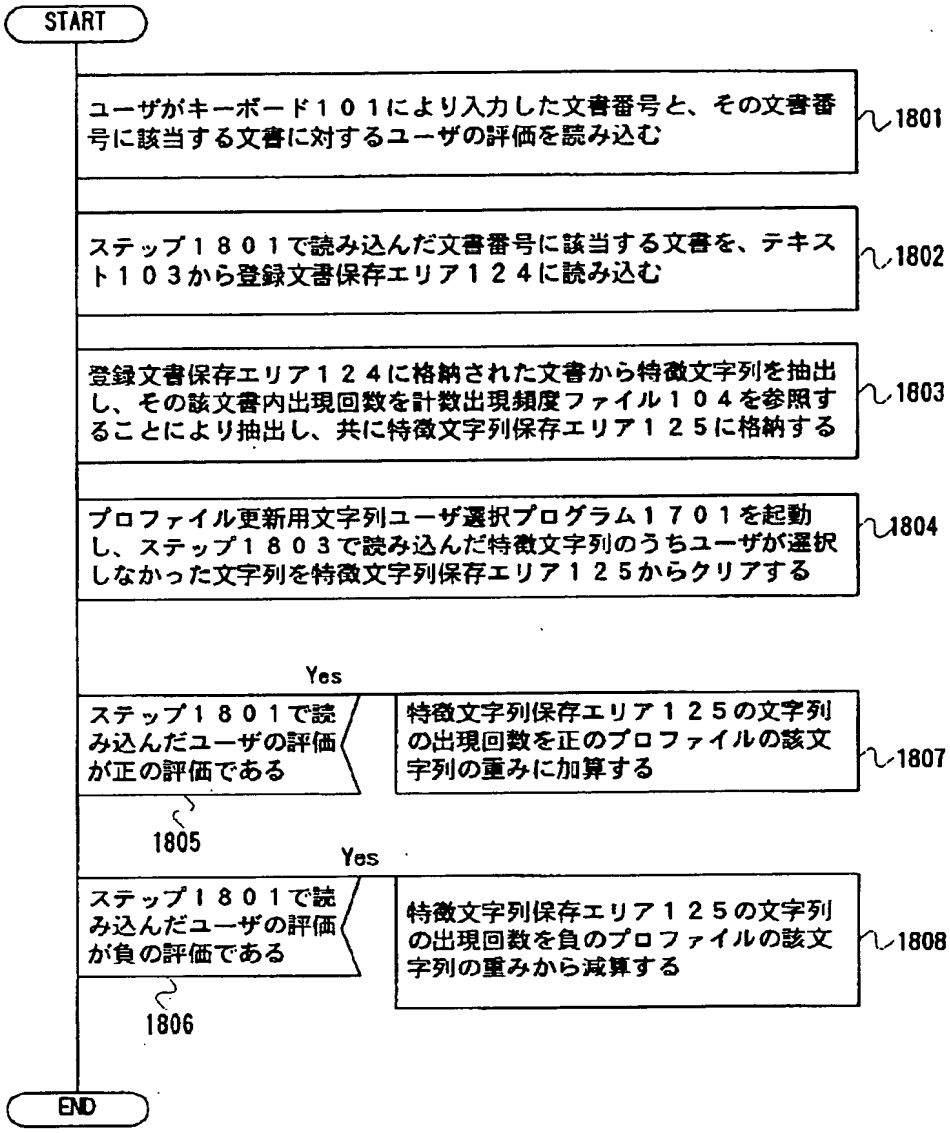
【図１６】

図１６


START
ユーザがキーボード１０１により入力した文書番号と、その文書番号に該当する文書に対するユーザの評価を読み込む
1301
ステップ１３０１で読み込んだ文書番号に該当する文書に出現する特徴文字列とその出現回数を、出現頻度ファイル１０４から特徴文字列保存エリア１２５に読み込む
1303
Yes
ステップ１３０１で読み込んだユーザの評価が正の評価である
特徴文字列保存エリア１２５内の文字列の出現回数を正のプロファイルの該文字列の重みに加算する
1306
1304
Yes （プロファイル更新用文字列選択ステップ）
ステップ１３０１で読み込んだユーザの評価が負の評価である
特徴文字列保存エリア１２５の文字列のうち、正のプロファイル１２０中の重みの上位のものに含まれるものを、特徴文字列保存エリア１２５からクリアする。
1601
1305
特徴文字列保存エリア１２５内の文字列の出現回数を負のプロファイルの該文字列の重みから減算する
1307
END

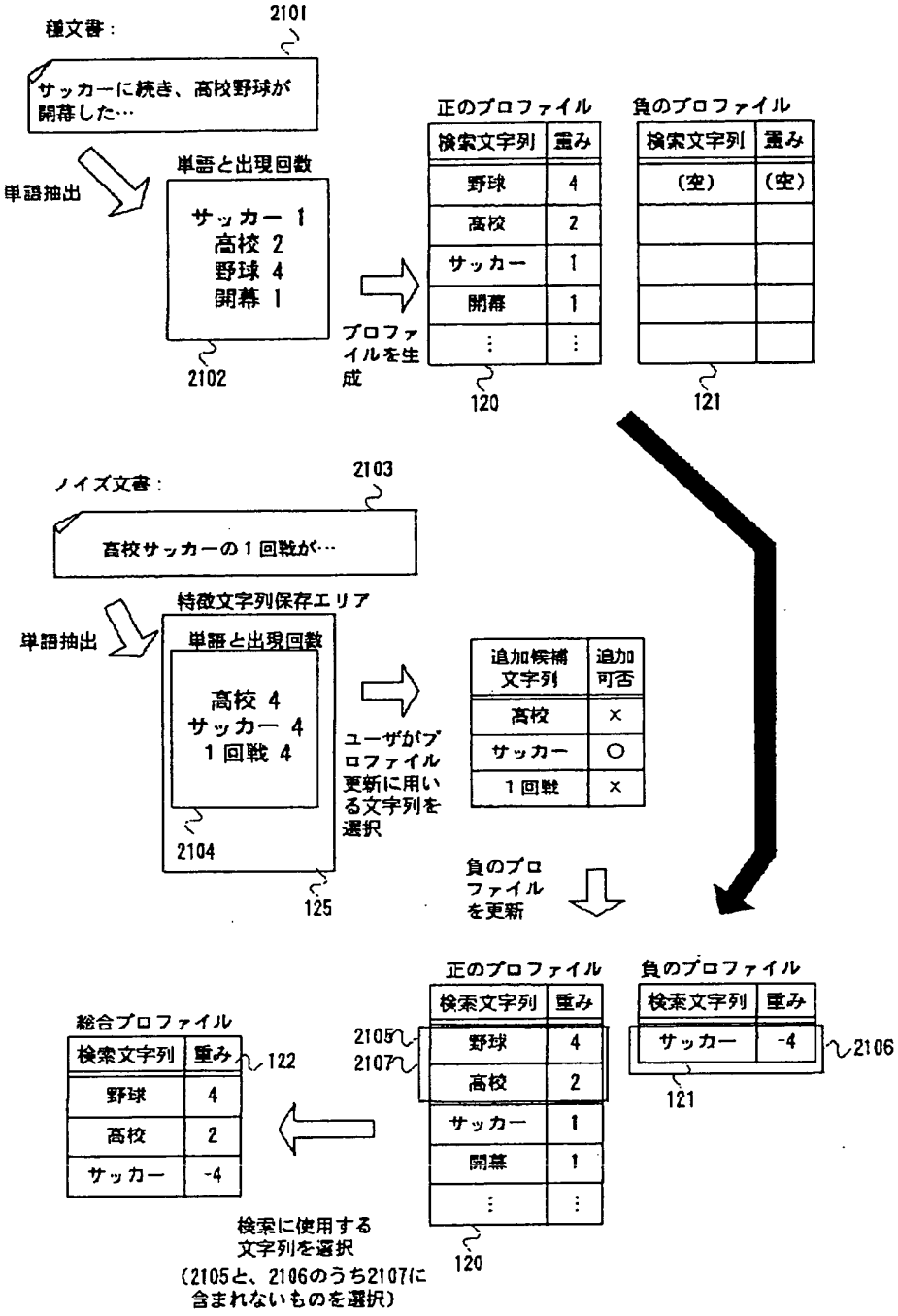
【図１８】

図１８


START
ユーザがキーボード１０１により入力した文書番号と、その文書番号に該当する文書に対するユーザの評価を読み込む
1801
ステップ１８０１で読み込んだ文書番号に該当する文書を、テキスト１０３から登録文書保存エリア１２４に読み込む
1802
登録文書保存エリア１２４に格納された文書から特徴文字列を抽出し、その該文書内出現回数を計数出現頻度ファイル１０４を参照することにより抽出し、共に特徴文字列保存エリア１２５に格納する
1803
プロファイル更新用文字列ユーザ選択プログラム１７０１を起動し、ステップ１８０３で読み込んだ特徴文字列のうちユーザが選択しなかった文字列を特徴文字列保存エリア１２５からクリアする
1804
Yes
ステップ１８０１で読み込んだユーザの評価が正の評価である
特徴文字列保存エリア１２５の文字列の出現回数を正のプロファイルの該文字列の重みに加算する
1807
1805
Yes
ステップ１８０１で読み込んだユーザの評価が負の評価である
特徴文字列保存エリア１２５の文字列の出現回数を負のプロファイルの該文字列の重みから減算する
1808
1806
END

【図２１】

図２１



フロントページの続き

(72)発明者　菅谷　奈津子
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所システム開発本部内

(72)発明者　松林　忠孝
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所システム開発本部内

(72)発明者　山口　明彦
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所システム開発本部内

(72)発明者　川下　靖司
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア事業部内

Ｆターム(参考)　5B075 ND03 NK02 NK32 PP30 PQ02
PQ40 PQ46 PR04 PR06 QM08
QS01 QS20 UU06